

nature

THE INTERNATIONAL WEEK OF SCIENCE



The rapidly
evolving
genome of the
seahorse PAGE 395

EVOLUTION AT A GALLOP

ECONOMICS

A POOR SENSE OF POVERTY

How the right statistics
can make a difference

PAGE 330

BIOTECHNOLOGY

UNNATURAL ENZYMES

Light-stimulated cofactor
transforms catalytic activity

PAGES 345 & 414

MATERIALS

LET'S GET CLINICAL

Polymers used in medicine
can be made safer by design

PAGE 335

NATURE.COM/NATURE

15 December 2016 \$10

Vol. 540, No. 7633



THIS WEEK

EDITORIALS

GENOMICS Sea horses give up some genetic secrets **p.316**

WORLD VIEW Academics must pick sides on populism **p.317**



MAPS Galaxy quest charts more than one billion stars **p.319**

A creeping corporate culture

The trend of turning universities into businesses is limiting research freedoms in traditionally liberal Scandinavian institutes. It is time for scientists to regain lost ground.

The current craze for all things *hygge* — a Danish word for general well-being, used by outsiders to sell everything from comforting cookbooks to comfortable cardigans — neatly summarizes the world's impression of Scandinavia.

From academia to civil society, many have long viewed this cluster of countries in northern Europe as a haven. The Nordic economic model, a combination of social market economy with lavish government expenditure, has guaranteed social peace, affordable higher education and — key for scientists — freedom to do research. Not for them the creeping trend of academic capitalism, with universities and researchers pressured by politicians to produce more than scholarly output.

But over the past decade, scientists in Scandinavia have seen corporate culture gaining ground. Researchers everywhere should be alarmed that the trend has spread so far — and academic leaders should take steps to regain some of the lost ground.

There was little *hygge* on show last month, for example, when the University of Copenhagen fired seismologist Hans Thybo, president of the European Geosciences Union. The official explanation for Thybo's dismissal — his alleged use of private e-mail for work, and telling a postdoc that it is legitimate to openly criticize university management — seems petty in the extreme. More than 1,000 scientists from across the world agree and have signed a petition, launched after *Nature* had reported about the case, to ask the university management to reverse its decision and reinstate Thybo. It should sincerely consider doing so. Denmark's leading university risks a nosedive in its global reputation if it is unable to produce more convincing grounds for parting ways with one of its best scientists. Thybo, an internationally esteemed geophysicist, experienced expedition leader and gifted teacher and supervisor, is held in high regard by colleagues, postdocs and students. But his unquestioned success in solid Earth science — funded, among other sources, by the oil and gas industry — failed to impress research managers who were apparently more concerned about his style and personal idiosyncrasies.

In many ways, the case symbolizes a growing power struggle and fuels an atmosphere of disconcertion and mistrust that was previously unknown in Denmark's academia. In 2003, the country introduced a law that requires the majority of members on university governing boards to be non-academics. Robbing researchers and students of the chance to decide on the matters they know best, the resulting hierarchical governing structure has raised real and repeated concerns over the freedom of research. Long-standing discontent has now surfaced, and the past couple of years have seen several scholars and scientists clashing head-on with business-orientated university managers.

It is of course desirable, and in the public interest, that universities produce innovation and jobs as well as beautiful science and lofty discourse. But corporate research managers must understand that the art of science starts with a tinge of intuition that is not easily commensurable with the textbook logic of managers and business administrators. Universities cannot exist without a degree of organization and prudent

allocation of finite material resources. But science depends on generous creative freedom and a sound measure of intellectual rebelliousness.

The Thybo case resonates because Danish science is already under pressure. In 2015, the country's newly elected centre-right government announced substantial cuts to funding for research and higher-education institutes. At the University of Copenhagen, more than 500 staff will lose their jobs. Harsh economic realities must not be an excuse

“Corporate culture is a threat to the profession and very pursuit of science.”

to get rid of respected researchers who fail to please corporate-minded managers. And for a cautionary tale of what can go wrong with the arrival of business thinking in academia, Danish policymakers need only look to their neighbours in Sweden. At the prestigious Karolinska Institute in Stockholm, the substitution of scientific leadership by a mostly

non-academic management has been blamed for mishandling cases of scientific fraud committed by medical researcher Paolo Macchiarini.

An obsession with accountability through metrics and excessive evaluation is already driving many scientists to distraction, not only in Scandinavia. But if corporate culture also means that scientists are fastidiously scrutinized in their every move, it is a threat to the profession and very pursuit of science.

As populism gains ground on both sides of the Atlantic (see page 317), courage and intellectual honesty are more valuable than ever. Corporate identity might work for a university as a marketing concept — but it offers little incentive for independent minds to speak out and make conclusions. In 1968, students across Europe forcefully revolted against conservative professors and academic institutions that had not changed for centuries. This time around, the threat is the colonization of universities by overzealous business types. ■

Protection at risk

Donald Trump's choice for head of the US environment agency is dismaying.

The bad news just keeps on coming. At the end of last week, reports surfaced of an intimidating exercise at the US Department of Energy, with staff asked to identify and name scientists and others who have attended UN climate conferences and helped to plan policy. Such a request — often associated with purges conducted by nascent authoritarians — bodes ill for honest intellectual inquiry at the department and in the wider US government.

There may be an innocent explanation. The e-mails and

correspondence requested might be used merely to inform and educate the incoming administration — but at this stage it is getting harder to give Donald Trump the benefit of what little doubt remains about the kind of US president he will be.

Opponents and critics of Trump — including many scientists — who were urged to judge him on his actions, rather than on his campaign rhetoric, are seeing their worst fears realized. Trump, they had been reassured, is a closet pragmatist whose pursuit of “the deal” will pull him towards the political centre.

Instead, as he builds his government, Trump is surrounding himself with like-minded ideologues who harbour extreme views on everything from national security and global warming to law enforcement, immigration and social policy. The Republican establishment that Trump bested during the primaries — already radicalized by the Tea Party and itself institutionally opposed to climate science — has fallen either in line or off the radar.

Trump's nomination of Oklahoma attorney general Scott Pruitt to lead the Environmental Protection Agency (EPA) is particularly worrisome. The EPA regulates pollution and chemicals at home, and it must play a powerful part in the United States' efforts to reduce greenhouse-gas emissions, which affect the entire planet. Yet Pruitt has won the affection of industry and Trump precisely because he has opposed such policies, time and time again.

Pruitt claims that the EPA has an activist agenda that threatens jobs and economic development. As attorney general, he challenged a federal rule intended to expand protections for waterways and wetlands. He fought a regulation that was designed to reduce the amount of mercury and other pollutants emitted by power-plant smokestacks. He was also among the state leaders who filed lawsuits against President Barack Obama's power-plant regulations. He disputed the landmark EPA judgment that climate change poses a danger to public health and welfare.

The content and language of these challenges focus on the legal tension between federal and state oversight. But written clearly between the lines is hostility to policies that dare to put the needs of the environment above the profits of industry.

Pruitt has demonstrated a wilful disregard for science, and has repeatedly put the interests of fossil-fuel companies ahead of those of his own constituents.

In at least one case documented by *The New York Times* last year, he used his office to help Devon Energy, an oil and gas company based in Oklahoma, challenge the EPA's estimates of methane emissions from natural-gas wells. Devon penned a letter in 2011, and Pruitt signed it and sent it to the agency. In a response to the *Times*, Pruitt acknowledged as much, but said it was the content of the letter, not the source, that mattered. “The oil and gas industry has been targeted unfairly by this administration,” Pruitt wrote. “The

“To make Scott Pruitt the head of the EPA would be a huge backward step.”

AG's office has particular interest in weighing in whenever any federal agency oversteps its authority to implement devastating policies.”

In fact, the US oil and gas industry has enjoyed an unparalleled resurgence during Obama's tenure. In large part, that's why natural gas is pushing coal out of the US electricity market, and why the price of oil has crashed. The spike in oil and gas activity is even causing earthquakes in Pruitt's home state. Nor is the EPA out of bounds in regulating methane emissions — that's the agency's job. More to the point, there were no policies to dispute, devastating or otherwise. Pruitt and the natural-gas industry just didn't like the data that were coming from EPA scientists, or its implications.

In 2012, he accused EPA officials of possible deception over measurements of methane emissions, and complained of a “wayward federal agency arbitrarily using unsubstantiated, inaccurate and flawed data to achieve a specific policy objective”. Pruitt has taken a similar attitude to climate science, which he has described as contested and uncertain.

To make him the head of the EPA would be a huge backward step, and one that should be opposed by scientists, policymakers and all who value the contribution of research to the public good. It is not yet a done deal. The US system demands that such appointments are confirmed by a vote in the Senate.

Senators must cast their votes — and make clear what they stand for — next year. Moderates across the political spectrum have a responsibility to make their voices heard to try to influence the outcome. This includes scientists and scientific organizations such as the National Academy of Sciences and the American Association for the Advancement of Science. The cause may seem forlorn, but it is not lost yet. ■

Symbolic sea horse

The genome sequence of this unusual creature offers clues to its unique traits.

The gods of Greek mythology were busy people. Poseidon, as well as having dominion over the sea and sending earthquakes, had a sideline in creating animals. His most celebrated design was the horse. Poseidon was so keen on his horses that he held onto some to pull his chariot through the waves. These first sea horses — called the hippocampi or, loosely, horse-monsters — had the tails of fish and two front hooves. They could be seen on a windy day, racing across the foam and waves of the sea's surface. That's why ocean breakers are still called white horses.

The sea horse, in other words — or its name at least — has a complicated origin story. Whereas Poseidon's mythical horses were considered the most beautiful creatures of the ancient world, the real sea horse has a tale of wonder of its own to tell. These fragile, elegant animals look like almost nothing else on Earth (except, naturally, a horse, and a distinctive part of the human brain). They are fish without scales and the usual fins. They are covered in bony plates. They swim upright. They form monogamous pairs. And most famously, the male sea horse experiences pregnancy — well, the closest that fish get to pregnancy — as he holds

and nurtures the developing embryo in a special pouch.

In a paper this week, scientists explore the bizarre features of the sea horse from the inside out (Q. Lin *et al. Nature* **540**, 395–399; 2016). They describe how they sequenced and analysed the genome of *Hippocampus comes*, the tiger tail sea horse (just to add to the morphological mix). The results offer some clues to the genetic basis of their unique traits.

A gene family with a role in embryo hatching shows high expression in the male brood pouch, the scientists say. And some potential regulatory elements are missing, which might help to explain the evolution of the sea horse's strange body shape. The animals eat through a tubular snout (no teeth) and, sure enough, the genome showed a lack of genes for enamel proteins, needed to make teeth. The absence of a gene called *tbx4*, a known regulator of limb development, may have contributed to the loss of pelvic fins. And to the unusual features of the sea horse we can add a relatively high evolutionary rate in their genes as compared to other fish.

As we gain understanding of what makes the sea horse so special, its future is far from assured. Many of the 46 or so known species are on the endangered list: drained from the sea as by-catch and sent around the world as live pets or as dried food and medicine. The sea horse is a powerful symbol, and one that has been used to catalyse conservation efforts, such as the creation of protected marine zones in places such as the Philippines. But pollution and habitat loss are also taking their toll — as they are on much of the wider ocean environment. The white horses may still skim across the surface, but the world of Poseidon is losing its magic. ■



Simply studying populism is no longer enough

Sociologist Matthijs Rooduijn explains why the darkening political mood must force academics to step up and choose sides.

Research on the political phenomenon of populism was traditionally a topic for historians. But in the past two or three decades, the academic field has grown to include political scientists, sociologists, communication scientists and psychologists. And we populism scholars have never been so popular. Since the US election, I have been inundated with requests from the media to talk about populism and why it seems to be catching on.

One of the most common questions is whether history is repeating itself — if the current situation resembles the political strife of the 1930s. I don't think so. This is new. Of course, there are some similarities, but there are also very large differences. Most importantly, fascists and national socialists are no populists, because they are not democratic. Populists are.

Right-wing politicians in the crop currently making headlines are populists in that they want the will of the people to be the point of departure for political decision-making. This 'general will' should, according to their populist message, be translated as directly as possible into actual political decisions. All institutions, rules and procedures that stand in the way of such a direct expression of the general will are conceived of as liabilities that should be removed as quickly as possible. Minority rights? They hamper the direct expression of the will of the people. Checks and balances? They delay the decision-making process. Political compromises? They lead to the dilution of policy proposals and therefore to a lack of decisiveness. Free media? It only represents the interests of the 'established order'.

A little bit of populism can act as a force for good by recognizing discontent and broadening the political agenda. But current right-wing populists go further: they infuse their populism with nativism, which argues that the nation is being threatened by 'dangerous others', such as immigrants or people of a non-majority race or religion. Populism and nativism are frequently confused and combined, but they are separate and distinct.

Initially I took the view that academics investigating these parties and politicians should approach their study as objectively as possible: they should try to be neutral observers who focus on understanding the causes and consequences of the rise of these political actors, without making moral judgements about the empirical patterns that they encounter.

As such, when I finished presentations on the causes and consequences of the rise of populist parties with an analysis of the relationship between populism and liberal democracy and the positive and negative sides of populism, my conclusion was always quite relaxed. In Europe, I used to say, we have strong liberal institutions, there is no all-pervasive populist zeitgeist and if populists manage to

make it into government it is usually as part of a junior coalition party.

However, things have changed. Populists in Hungary and Poland seriously challenge liberal institutions, populist discourse has become more widespread and, when in government, populists are no longer merely junior partners.

Most disturbingly, mainstream parties in Europe seem to have incorporated elements of populism's illiberalism. In France, for instance, the enduring state of emergency established after last year's terrorist attacks in Paris has led to abusive raids and infringements of people's rights. Many mainstream parties in Western European countries are choosing security over liberty — probably because they feel the radical-right populists breathing down their necks.

So I have changed my mind and my approach. I will remain as neutral as possible in my academic work, but I increasingly feel obliged to take part in the public debate about this topic, and to warn in the media of the increasing tension between populism and liberal democracy.

More academics must speak out and warn about where we are heading. Part of this is immediate self-interest. There is no reason to expect that academia will be immune to the kind of populist interferences that we are now seeing in Hungary and Poland. Populist attacks on checks and balances and media freedom might well spill over into attacks on academia as well. After all, populists not only attack political and economic elites; they also target 'snobby intellectuals' in academia. In fact, such attacks on academics are happening in Turkey right now.

Academics also have a moral obligation to protect liberal democracy. By promoting social and political pluralism, the system produces the circumstances under which researchers can do their jobs and science can flourish. Researchers depend on it.

Events this year have been worrying. And the first big test of 2017 comes uncomfortably close to home for me. The populist Geert Wilders of the Dutch Party for Freedom is leading the opinion polls in the build-up to the national elections in March. He might well win, but it's highly unlikely that he will become the next prime minister. He will have to form a government coalition if he does not get more than 50% of the seats. Most other parties have already ruled out collaboration with him, so I think it is very unlikely that he will govern. However, with four or more mainstream parties forming a coalition, Wilders's message that the political establishment is colourless and all the same might become even more popular. ■

*Matthijs Rooduijn is an assistant professor in the department of sociology at Utrecht University in the Netherlands.
e-mail: m.rooduijn@uu.nl*

**ACADEMICS HAVE A
MORAL
OBLIGATION
TO
PROTECT
LIBERAL
DEMOCRACY.**

NEUROSCIENCE

Transplanted brain cells calm fear

Mice that receive neuron transplants are better at forgetting fearful memories than those without transplants.

Yong-Chun Yu at Fudan University in Shanghai, China, and his colleagues studied mice that had learned a fearful memory and were then trained to forget it. After this 'extinction' training, fear memories often come back spontaneously with time or in response to a stimulus. But the team found that this later recurrence was reduced when embryonic neurons that make a neurotransmitter called GABA were transplanted into the animals' brains two weeks before the extinction training.

The neurons were transplanted into the amygdala, a brain region associated with fear, and the findings suggest that the cells may have returned it to a more pliable, juvenile state. This could increase the effectiveness of fear-extinction training, the authors suggest.

Neuron <http://doi.org/bvp3> (2016)

ASTROPHYSICS

Dark matter may not be so clumpy

An analysis of almost 15 million distant galaxies reveals that dark matter may be slightly less dense and more evenly distributed throughout space than was thought.

Dark matter makes up one-quarter of the Universe's mass, but is invisible and its presence can only be inferred from its gravitational effects. A team led by Hendrik Hildebrandt of the Argelander Institute for Astronomy in Bonn, Germany, and Massimo Viola of Leiden University in the Netherlands

examined galaxy images taken by the European Southern Observatory's VLT Survey Telescope in Chile as part of the Kilo-Degree Survey. The researchers measured cosmic shear: the distortion of the shapes of background galaxies due to light that is warped by the gravitational effects of large-scale structures such as galaxy clusters. The team statistically measured how dark matter subtly distorted the galaxy images, and inferred its density from this.

If future measurements confirm this more-even distribution of dark matter, astrophysicists might need to

revise their models of how the Universe evolved.

Mon. Not. R. Astron. Soc. (in the press); preprint at <https://arxiv.org/abs/1606.05338> (2016)

BIOMATERIALS

How additives preserve vaccines

Scientists have found additives that, at low concentrations, extend the life of vaccines at room temperature.

High levels of sugar stabilize virus particles in vaccines, but the mechanism was unclear. Francesco Stellacci at the Swiss Federal Institute of Technology

in Lausanne and his colleagues studied how sucrose and two other additive candidates affect viruses over time. They found that low concentrations of the polymer polyethylene glycol and gold nanoparticles mimicked the effects of sugar, increasing the half-life of a virus called adenovirus type 5 from 7 days to more than 30 days at room temperature.

The team concludes that high levels of sugar keep viruses structurally intact mainly by making the vaccines more viscous. For the other additives, particles act directly on the virus's protein shell to prevent it from degrading. The findings



ENERGY

Solar power pays off

Solar-cell production generates high levels of greenhouse-gas emissions, leading some to question the environmental sustainability of the booming business (pictured) — a concern now allayed by scientists in the Netherlands.

A team led by Atse Louwen at Utrecht University studied developments in photovoltaic production around the world between 1976 and 2014. The authors found that, thanks to ongoing improvements to the technology and

production methods, every doubling of global photovoltaic capacity was associated with a drop of up to 13% in the energy used during system production, and a fall of as much as 24% in greenhouse-gas emissions.

Even in the worst-case scenario, whereby solar panels perform at their lowest efficiency levels, the industry is set to break even in 2017 in terms of energy use and in 2018 for emissions.

Nature Commun. 7, 13728 (2016)

W. T. FITCH ET AL.

should aid in the design of better additives, which could reduce the high cost of keeping vaccines cold to maintain their potency during distribution.

Nature Commun. 7, 13520 (2016)

MATERIALS

Graphene putty feels the beat

A dash of graphene can transform the stretchy material known as Silly Putty into a pressure sensor that can monitor a human pulse and even the steps of a small spider.

Jonathan Coleman at Trinity College Dublin and his colleagues mixed graphene flakes — consisting of roughly 20 layers of carbon atoms, and measuring up to 800 nanometres in length — with homemade Silly Putty, a cross-linked silicone polymer. This produced a material, dubbed G-putty, that conducted electricity. Its resistance changed markedly when the authors applied the slightest pressure, making it more than ten times more sensitive than typical pressure sensors.

The team used G-putty to take accurate blood-pressure measurements and record the steps of a 20-milligram spider. *Science* 354, 1257–1260 (2016)

ANIMAL BEHAVIOUR

Ants 'talk' by swapping spit

Oral fluid exchanged between ants contains molecules that the insects might use to communicate.

Ants were generally thought to share only nutrients and enzymes through a mouth-to-mouth feeding process called trophallaxis. But when Adria LeBoeuf at the University of Lausanne in Switzerland and her co-workers analysed the oral liquid of the species *Camponotus floridanus*, they found 64 microRNAs, 49 long-chained hydrocarbons, a hormone that regulates growth and more than 50 proteins

involved in development, digestion and immunity.

The hydrocarbons could contribute to a characteristic colony odour, and the hormone may influence larval growth and development. When the team added the hormone to the food of worker ants, more than twice as many of the larvae they reared reached adulthood, compared with those that were not exposed. The findings suggest that trophallaxis facilitates communication and helps the colony to develop, the authors say.

eLife 5, e20375 (2016)

ASTRONOMY

Gaia charts one billion stars

The positions of more than one billion stars in our Galaxy have been mapped with unprecedented precision by the European Space Agency's Gaia satellite (artist's impression **pictured**).

The craft launched in 2013 with the aim of making the most detailed ever 3D map of a portion of the Milky Way — including distances to stars from Earth, which are difficult to measure. Lennart Lindegren at Lund University in Sweden and his colleagues processed the first version of the data set, whose uncertainties are one-third of the size of those from the satellite's predecessor, Hipparcos. The catalogue currently includes some two million measurements of parallax — a star's apparent shift in position in the sky as Earth orbits the Sun — which allows scientists to determine the star's distance from Earth.

The catalogue should eventually allow researchers

to improve on estimates of the locations of most distant galaxies and the expansion of the Universe, which are measured relative to distances between nearer objects.

Astron. Astrophys. 595, A4 (2016)

SYNTHETIC BIOLOGY

Designer cells treat diabetic mice

Kidney cells grown in the lab have been engineered to both sense and quickly respond to changes in blood glucose levels.

In diabetes, cells in the pancreas called β cells are either absent or do not produce the correct amount of insulin to regulate blood glucose levels. Jörg Stelling and Martin Fussenegger at the Swiss Federal Institute of Technology Zurich in Basel and their colleagues genetically modified human kidney cells to enable the cells to detect blood glucose levels and produce an appropriate amount of either insulin or another hormone called GLP-1, which stimulates insulin production. The cells brought blood sugar levels down to normal when implanted into mice with type 1 or type 2 diabetes, without notable adverse effects.

In mice with type 1 diabetes, the 'designer' cells were more efficient at restoring normal glucose levels after a three-week period than implanted β cells.

Science 354, 1296–1301 (2016)

PRIMATOLOGY

Macaques vocally equipped to speak

Monkeys could talk, if they only had the right brain circuitry.

An influential 1969 paper examined the cadaver of a rhesus macaque (*Macaca mulatta*) and concluded that its vocal anatomy was not capable of speech. To reassess this claim, a team led by Tecumseh Fitch at the University of Vienna and Asif Ghazanfar at Princeton University in New Jersey



X-rayed live long-tailed macaques (*Macaca fascicularis*, **pictured**) as they made various sounds, such as threat calls.

Using the scans, the authors developed a computer model of the macaque vocal tract. This suggested that the monkeys do have the anatomy to make speech sounds, including five vowels and even the phrase "Will you marry me?"

Monkeys can't speak because they lack the brain circuitry required for fine motor control, vocal learning and other attributes necessary for speech, the authors say.

Sci. Adv. 2, e1600723 (2016)

ECOLOGY

Extinctions on the warm front

Hundreds of species are not adapting quickly enough to cope with global warming, and are disappearing from local areas in the warmest parts of their ranges.

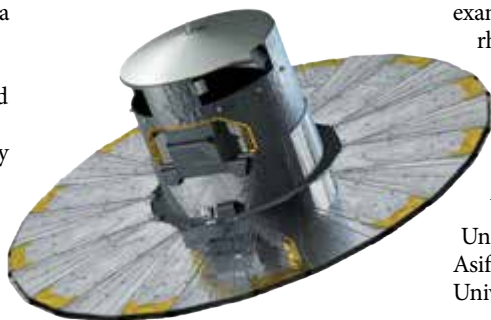
John Wiens at the University of Arizona in Tucson analysed 27 studies on 976 species, many of which have shifted their range in response to climate change. In almost half, populations have disappeared from the hottest edges of their ranges as the global climate has warmed. These local extinctions were more common in the tropics and subtropics than in other climates, in fresh water than in the sea or on land, and in animals than in plants.

PLoS Biol. 14, e2001104 (2016)

► **NATURE.COM**

For the latest research published by Nature visit:

www.nature.com/latestresearch



D. DUCROS/ESA

SEVEN DAYS

The news in brief

POLICY

Catch commitment

Some of the world's largest seafood companies have committed to clean up their industry in a statement issued on 14 December. The statement is the result of a process started by scientists at the Stockholm Resilience Centre, who in 2014 identified that 13 "keystone" fisheries companies controlled 11–16% of all wild marine catches. Eight of these companies have now agreed to improve their transparency and the traceability of their fish, and to "engage in science-based efforts to improve fisheries".

Habitat laws

The European Union has decided against overhauling major pieces of legislation that protect birds and natural habitats. Conservationists celebrated after the findings of a review, announced on 7 December, stated that the laws were "fit for purpose" and would not be opened up to reforms that could weaken them. The birds and habitats directives protect more than 1,000 species and 1 million square kilometres of land in the EU. On 8 December, the European Commission announced that it was taking France to court for breaching the birds directive by failing to protect wild species.

Peatland protection

Indonesian president Joko Widodo announced a moratorium on 5 December on development activities that damage the nation's vast peatlands. The action will, among other things, prevent the conversion of peatland into oil-palm plantations. It comes after catastrophic fires linked to the clearing choked the region's air with smoke last year, causing health problems, as well as an estimated



First US astronaut to orbit Earth dies

John Glenn, the first US astronaut to orbit Earth and an icon of the US space age, died in Columbus, Ohio, on 8 December, aged 95. Trained as a Marine Corps pilot, in February 1962 Glenn became the second person to circle the planet in space, after the Soviet Union's Yuri Gagarin. Glenn completed 3 laps over 5 hours

in NASA's *Friendship 7* capsule (pictured). He later entered politics and served as Democratic senator for Ohio for 24 years, working on issues including energy policy and nuclear non-proliferation. In 1998, aged 77, Glenn flew aboard the space shuttle *Discovery* as the oldest astronaut ever to do so.

US\$16 billion of economic damage. Indonesia has pledged to reduce its carbon emissions — the bulk of which come from deforestation and peatland destruction — by 29% by 2030, compared with projected levels.

SPACE

Methane problem

India's Mars Orbiter Mission (MOM), which made the country's space agency only the fourth to successfully send a probe to the red planet, has a problem with its methane sensor, according to online news outlet Seeker. Measurements of atmospheric methane by MOM had been eagerly awaited, but no such data have been released since the probe reached Mars in

September 2014. A methane specialist at NASA told Seeker that although the sensor collects measurements, a design flaw means that it does not process and send back spectroscopic data in a usable form. The Indian Space Research Organisation, which has not acknowledged the problem, will repurpose the sensor into an albedo mapper, says the NASA scientist.

TECHNOLOGY

AI research

Breaking with its usual secretive approach, computer giant Apple announced on 6 December that it will, for the first time, allow its artificial-intelligence (AI) researchers to publish their work. Critics

have said that prohibiting researchers from engaging with the AI community was part of the reason that the company had fallen behind in the field. Meanwhile, Uber, the car-hailing company, announced the previous day the creation of Uber AI Labs in San Francisco, California, in a bid to improve its driverless-car technologies, among other things.

PEOPLE

NIH clinical chief

The US National Institutes of Health (NIH) has chosen a retired army major general to head its troubled Clinical Center in Bethesda, Maryland. On 9 December, the agency announced that James Gilman, a cardiologist,

SERGEI ILITSKY/EPA

will join the centre as its first chief executive. In 2015, federal inspectors found widespread contamination in a facility that manufactures experimental drugs and other medical products for the centre. John Gallin, the centre's director at the time, stepped down and has now been appointed its chief scientific officer.

Trump transition

US president-elect Donald Trump has made a number of key cabinet nominations. On 13 December, he nominated Rex Tillerson, the chief executive of oil giant Exxon Mobil, to be his secretary of state. As the United States' top diplomat, Tillerson (pictured) would have a prominent role in climate policy — such as in negotiating a US exit from the 2015 Paris climate accord, something Trump pledged during his campaign. On 7 December, Trump picked Oklahoma attorney general Scott Pruitt to lead the Environmental Protection Agency. Pruitt has questioned the science underlying global warming, and is one of dozens of state officials who have mounted a legal challenge to President Barack Obama's limits on carbon emissions from power plants. And as *Nature* went to press, Trump was expected to announce Cathy McMorris Rodgers



to lead the Department of the Interior, which oversees federal public lands and natural resources. McMorris Rodgers, a congresswoman from Washington state, has also expressed doubt over human-induced climate change and has advocated expanding oil and gas development. All nominations will need approval from the Senate. See page 315 for more.

PUBLISHING

Anonymity ruling

PubPeer, a website that allows anonymous reviews of scientific papers, has won a key legal battle against a cancer researcher who claims that defamatory remarks on the site cost him a job. In a ruling published on 6 December, judges in a Michigan appeals court reversed a 2015 decision that mandated the site to reveal the identity of anonymous commenters after the scientist, Fazlul Sarkar, sued them. Sarkar can continue to pursue

a defamation case, judges said, but he is not entitled to reveal the identities of PubPeer commenters, whose anonymity is protected by the US First Amendment.

Impact-factor rival

Publishing giant Elsevier launched the CiteScore index on 8 December — a rival to the Journal Impact Factor (JIF), one of science's most contentious metrics. CiteScore ranks journals using a similar formula to that of the JIF, but it covers twice as many journals and includes tweaks that produce some notably different results — including lower scores for some high-JIF journals. See page 325 for more.

EVENTS

Chinese fraud cases

The Natural Science Foundation of China (NSFC) has released a list of 61 cases of scientific misconduct discovered during 2015 and 2016 in research that it had funded. The cases involve plagiarism, falsified data and images, authorship problems, and faked publication lists in grant applications. Many of the researchers involved were caught using fake peer reviewers. The list, which was posted on the agency's website after a press conference in Beijing on 12 December, is part of the country's ongoing

crackdown on research misconduct. Punishments meted out to those who received the grants — and, in the case of retracted papers, to first and corresponding authors — include publicly criticizing the researchers, revoking their grants, recovering funds and banning them from applying for grants from the NSFC for up to seven years.

FUNDING

Climate coalition

Bill Gates on 12 December announced the launch of an ambitious effort to commercialize emerging low-carbon technologies in industry, transport, agriculture and the energy sector. The Breakthrough Energy Ventures fund will be “guided by science” and led by an alliance of 20 of the world's richest people. They aim to invest more than US\$1 billion over the next 20 years in climate-friendly technologies including clean power generation and electricity storage. Investors contributing to the fund, which will begin next year, also include Alibaba founder Jack Ma and Amazon's Jeff Bezos.

R&D funding slips

Spending on research and development (R&D) by governments and higher-education institutions in the now 35 member states of the Organisation for Economic Co-operation and Development (OECD) fell in 2014 for the first time since the organization began collecting the data in 1981. The *OECD Science, Technology and Innovation Outlook 2016*, published on 8 December, also shows that the share of public R&D in total government spending fell between 2000 and 2015 in seven of the ten leading member countries. The three exceptions are Korea, Germany and Japan.

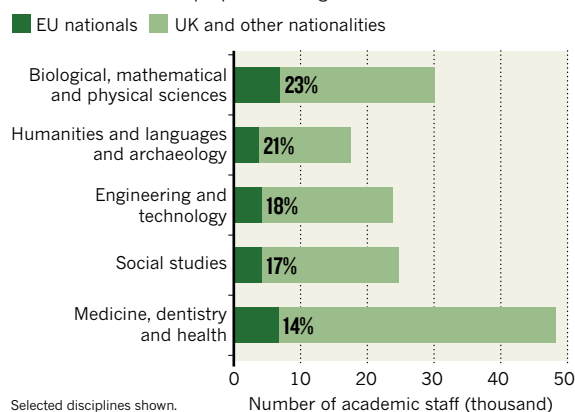
SOURCE: HESA

TREND WATCH

A UK parliamentary inquiry last week published evidence on how the country's higher-education system might be affected as a result of Brexit, the split from the European Union. More than 31,000 academics at UK universities are non-British EU citizens and may lose their right to live in the United Kingdom after Brexit. Statistics sent to the inquiry by the UK Department for Education show that these individuals are concentrated in the sciences. See go.nature.com/2hsxra3 for more.

HOW BREXIT COULD AFFECT SCIENCE WORKFORCE

Non-UK EU staff make up around 16% of academics at UK universities — but the proportion is higher in the sciences.



NEWS IN FOCUS

POLITICS Departing Italian prime minister interfered in academic affairs **p.324**

BIBLIOMETRICS Publishing giant launches new measure of journal impact **p.325**

POLICY Does it matter whether Donald Trump has a science adviser? **p.327**



DEVELOPMENT What researchers know about the world is wrong **p.330**

MATTHEW TOSTEVIN/REUTERS



The *Minecraft* video game is popular with children; now a version is being used to test artificial-intelligence programs.

TECHNOLOGY

Virtual worlds open doors to bevy of AI programs

Artificial-intelligence algorithms can learn a lot from playing immersive 3D video games.

BY DAVIDE CASTELVECCHI

The *Minecraft* video game was familiar to José Hernández-Orallo long before he started using it for his own research. The computer scientist, who devises ways to benchmark machine intelligence at the Polytechnic University of Valencia in Spain, first watched his own children play inside the 3D virtual world, which focuses on solving

problems rather than shooting monsters.

In 2014, Microsoft bought *Minecraft*, and its science arm, Microsoft Research, gave its own researchers access to a new version of the game that allowed computer programs, as well as people, to explore and customize the 3D environment. Then, after inviting a small group of outside researchers that included Hernández-Orallo to download the machine-friendly version of the world,

last July, Microsoft made it freely available to anyone, with the goal of speeding up progress in artificial intelligence (AI).

Now other companies have followed suit. On 3 December, DeepMind, a unit of Google headquartered in London, opened up its own 3D virtual world, DeepMind Lab, for download and customization by outside developers. The company initially created the world to train its own AI programs. Two days later, OpenAI, ►

► a research company in San Francisco, California, co-founded by entrepreneur Elon Musk, released a 'meta-platform' that enables AI programs to easily interact with dozens of 3D games originally designed for humans, as well as with some web browsers and smartphone apps.

All three releases provide researchers and software developers with easy ways to test programs in previously unseen situations, and for the programs to acquire new skills by teaching themselves to navigate novel situations that resemble real-world scenarios. "Environments like these have a very important role to play in the future of AI," says Pedro Domingos, a machine-learning researcher at the University of Washington in Seattle.

ATARI ALGORITHM

Games have been test beds for AI for decades, but, typically, the algorithms have played them following predefined strategies. In recent years, the focus has shifted to machines that could learn from their own experience. In early 2015, DeepMind unveiled an algorithm that taught itself how to play classic Atari arcade games better than any human, by trial and error, without being told the goals of the games.

Such games are simple 2D worlds, though. 'First-person' 3D video games such as *Minecraft* — which visually embed the player in the environment — are a much closer approximation to the real world, and so make more sophisticated test beds.

Minecraft enables users to interact with virtual bricks, and use them to build structures,

in addition to navigating and interacting with predefined structures. The version now available to developers, called Malmö, allows algorithms to do the same. Hernández-Orallo, for example, is using this to explore whether the environment can be used to create benchmarks for machine intelligence. Algorithms could compete to arrange bricks into something that looks the most like a certain object, say, or to navigate a maze — testing a much wider range of skills than the Turing test, the most famous test of machine intelligence, which focuses on the ability of an AI to chat like a human.

One of the things that made *Minecraft* attractive for conversion into an AI test bed is that it already enabled players to communicate using text messages. This could help an AI to learn to collaborate with humans in the real world, says computer scientist Katja Hofmann of Microsoft Research in Cambridge, UK, who led the team that created Malmö.

ROBOT REHEARSAL

Virtual worlds are also particularly useful for developing AIs that are destined to eventually operate as physical robots, says Hofmann, because such environments are cheaper to customize, and faster and safer to practise in than the real world. They also allow robotics researchers to focus purely on the intelligence part of the equation — the mechanical challenges of physical robots can be a distraction.

In addition to Hernández-Orallo, Microsoft Research has collaborations with a handful of

research labs that are using Malmö projects. But Hofmann suspects that many more are using it, perhaps around 100.

DeepMind Lab similarly allows researchers to create structures such as mazes, and their algorithms can learn to collect rewards as well as to navigate. DeepMind has also been experimenting with integrating "more naturalistic elements", such as undulating terrains and plants, into the platform, says a spokeswoman. Now that the environment is open, the company hopes that other researchers will help to make the environments more challenging for the algorithms. "By open-sourcing it, we are allowing the wider research community to get involved in shaping this," she says.

OpenAI's meta-platform, Universe, takes things even further. By providing multiple, radically different environments for the same AI to sample, it could help to attack one of the hardest problems in the field: creating algorithms that can use previous experience when faced with new situations. For instance, deep neural networks, which mimic the layers of brain cells in the visual cortex, can quite quickly learn to navigate a 3D maze, but cannot transfer the knowledge to navigate another maze. "If you change the colour of the maze, the system is completely lost," says Hernández-Orallo. "State-of-the-art technology fails dramatically."

Microsoft is now working to make Malmö available through Universe. "Having a community platform will accelerate everyone," says Greg Brockman, co-founder and chief technology officer of OpenAI. ■

POLITICS

No fond farewell for Italy's premier

Scientists feel let down by ex-Prime Minister Matteo Renzi.

BY ALISON ABBOTT

Italian politics is in turmoil after the resignation of Prime Minister Matteo Renzi — but researchers say that they are not particularly sad to see him go.

In his almost three years in charge, Renzi promised improvements for universities and science but failed to raise the status of research in the country, according to scientists who complain that he also directly interfered in academic affairs.

"Renzi became prime minister at a time of serious economic and social crisis, and he

injected a sense of energy and optimism into the university and research sector," says biologist Cesare Montecucco of the University of Padua. "Our expectations were raised, but they were mostly disappointed."

Renzi resigned on 7 December, three days after constitutional reforms that he proposed were defeated in a referendum. He stayed on to push through a 2017 budget that sees no significant increase for Italy's chronically underfunded university and research system. (Exact figures for research spending have not yet been released.) Italy's research and university funding per head is among the lowest in Europe

— although the country does produce a greater share of highly cited research papers than the European Union average. Little has changed on that score during Renzi's tenure, say Montecucco and other scientists.

Renzi has not delivered what they have long campaigned for: less bureaucracy for research institutions and a new research-grants agency along the lines of the US National Science Foundation.

FUNDING FALLACIES

Most controversial has been Renzi's November 2015 decree creating a €1.5-billion (US\$1.7-billion) centre for genomics in Milan. Known as the Human Technopole, it will focus particularly on personalized medicine and nutrition. The country's 2017 budget foresees annual funding of well over €100 million, beginning in 2018.

Although some are grateful for the research funding, many scientists have complained that this major investment in a single new project is inappropriate when most other public research institutes are starving for cash. They also strongly objected to the fact that it was planned by Renzi with a few

METRICS

chosen scientists, behind closed doors.

In September 2016, Renzi floated the idea of creating 500 elite professorships known as Natta chairs (after Italian chemist and Nobel laureate Giulio Natta), to be awarded mainly to Italians working abroad. They would be selected by 25 evaluation panels whose chairs the prime minister would nominate. Thousands of academics signed an open letter in October complaining that Renzi had designed the programme without discussing it with universities. The letter also protested against the involvement of politics in the selection.

Regulations for the Natta selection procedure have not yet been published, and so scientists hope that the next government will ensure

“This means that weaker universities in the south will lose even more money.”

that the process remains inside the academic community.

“Nomination of panel chairs by the prime minister is just not acceptable,” says physicist Giorgio Parisi of the University of Rome La Sapienza, a prominent critic of the process. “It is a political choice to do the selection independently of Italian universities, but then you could turn to external academic organizations, like Europe’s national academies.”

BUDGETARY BLUES

Parisi is also unhappy with aspects of the 2017 universities budget. In particular, €271 million will now be reallocated to the university departments that are judged by the national evaluation agency ANVUR to have the best research performance. Parisi thinks that rewards for high performers should come from new money, rather than being transferred from a general university budget that is already stretched thin. “This government reallocation means that weaker universities in the south will lose even more money, and this would be a social disaster,” he says.

An interim government will hold down the fort until new elections are held, which could take place next year. Uncertainty is set to continue. Populist and protest parties, particularly the Five Star Movement led by comedian Beppe Grillo, are likely to make substantial gains in the next election.

These parties do not have strong scientific agendas. Italian senator-for-life Elena Cattaneo, who is also a neuroscientist at the University of Milan, is taking a wait-and-see perspective. “One or two populists in the current parliament have shown themselves to be more open to discussion on scientific topics than members of mainstream parties,” she says. ■

Impact factor gets heavyweight rival

CiteScore uses larger database and gets different results.

BY RICHARD VAN NOORDEN

One of science’s most contentious metrics has a flashy new rival. On 8 December, publishing giant Elsevier launched the CiteScore index to assess the quality of academic journals.

Although the index ranks journals with a formula that largely mimics the influential Journal Impact Factor (JIF), it covers twice as many journals — 22,000 to the JIF’s 11,000 — and its formula includes tweaks that produce some notably different results. These include lower scores for some high-JIF journals (see ‘A new measure of journal impact?’).

If CiteScore becomes popular, these quirks could change the behaviour of journals hoping to maximize their score, say analysts. But CiteScore’s debut comes at a challenging time for such metrics. It’s not obvious that there is an appetite for a competitor to the JIF, and scientists note that no matter what differences CiteScore provides, it will have to survive the same criticisms that are lobbed at its rival — most notably that the JIF is so commonly promoted by publishers as a yardstick for ‘quality’ that researchers are judged by the impact factor of the journal in which their work appears, rather than by what they actually write.

“In my view, journal metrics should always be accompanied by health warnings that are at least as prominent as the ones you see

on cigarette packets,” says Stephen Curry, a structural biologist at Imperial College London. “Such metrics are at the root of many of the current evils in research assessment.”

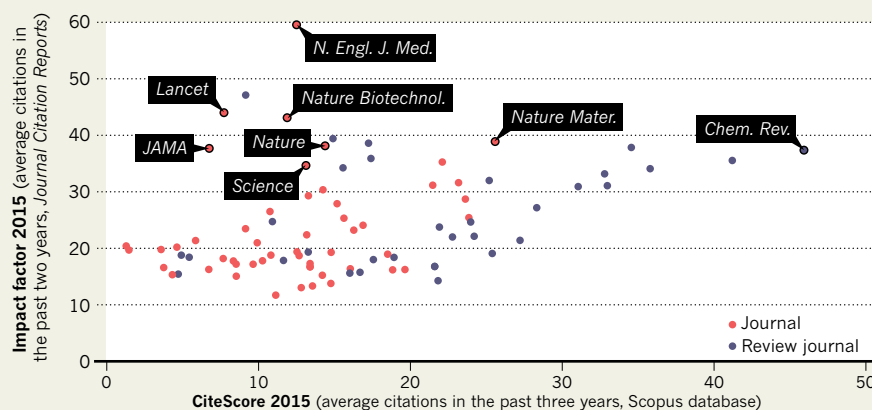
Amsterdam-based Elsevier has for many years provided a suite of analytical indicators, including journal metrics that have never become as popular as the JIF. It says that it has launched CiteScore owing to “overwhelming demand” from authors and editors.

The publisher is uniquely placed to challenge the JIF’s hegemony. It owns the Scopus database, a record of article abstracts and their reference lists. Aside from Web of Science, on which the JIF is based, it is the world’s only reasonably comprehensive and carefully curated citation database. But Scopus is bigger, enabling scientists, librarians and funders to check the popularity of many more journals. Furthermore, unlike the JIF, which is available only to subscribers, CiteScore figures will be free online for anyone to view and analyse, although full details of the documents included in the calculations are visible only to subscribers.

When it comes to their underlying formulae, CiteScore and JIF are near-doppelgängers. To score any journal in any given year, both tot up the citations received to documents that were published in previous years, and divide that by the total number of documents. The most popular version of the JIF looks at research articles published in the previous ▶

A NEW MEASURE OF JOURNAL IMPACT?

Journals that have a high impact factor, a measure of the average number of citations that their articles receive, don’t necessarily score so well on a new indicator, CiteScore. The latest metric includes documents such as editorials, letters and news items, which attract fewer scholarly citations.



The highest-scoring journal on both measures is CA: A Cancer Journal for Clinicians with an impact factor of 138 and CiteScore 66.

SOURCE: JCR/CITESCORE METRICS

► two years, whereas CiteScore counts the previous three.

But one significant difference leads some high-JIF journals, such as *Nature*, *Science* and *The Lancet*, to do worse in CiteScore. The new metric counts all documents as potentially citable, including editorials, letters to the editor, corrections and news items. These are less cited by scholars, so they drag down the average. *The Lancet*, for instance, drops from a healthy average of 44 in JIF — putting it in 4th position — to 7.7 in CiteScore, outside the top 200.

Such a distinction could have major consequences for the behaviour of publishers. “As there is intense competition among top-tier journals, adoption of CiteScore will push editors to stop publishing non-research documents, or shunting them into a marginal publication or their society website,” predicts Phil Davis, a publishing consultant in Ithaca, New York.

NUANCED CONTENT

The Lancet, *Nature* and other journals declined to comment on CiteScore. But Jeremy Berg, the editor-in-chief of *Science*, says that the journal is “very proud of our content that lies outside traditional research reports and articles” and that “any metric that is based on citation data alone will undervalue the impact of such non-research content”.

“The portfolio performance of all publishers may look a bit different using CiteScore metrics, including Elsevier, but all publishers gain in that they can explore the performance of more of their titles because of the broader coverage of Scopus,” says Lisa Colledge, director of research metrics at Elsevier. She says that CiteScore should be used only to compare related journals, not to compare raw scores across different fields. For example, the index ranks *The Lancet* 25th out of 1,549 ‘general medicine’ journals — putting it in the top 98th percentile of journals in that subject category.

Clarivate Analytics in Philadelphia, Pennsylvania, which bought the JIF and the Web of Science this year from Thomson Reuters, says that it doesn’t see any new insights in CiteScore. Other, more complex metrics — including several published by Elsevier and Thomson Reuters — have been developed to rank journals in the past, but none has yet proved as popular as the JIF. “If anything, another, different metric will reinforce the status that the JIF has as the definitive assessment of journal impact,” says Clarivate spokesperson Heidi Siegel.

Some even wonder whether Elsevier, which publishes more than 2,500 journals, should be producing CiteScore at all. The JIF has always been owned by non-publishers. “I question the appropriateness of a publisher getting involved with the metrics that evaluate the very content that it publishes,” says Joseph Esposito, a publishing consultant in New York City. But Elsevier says that it is “a provider of information solutions as well as a publisher,” and treats all the publishers it analyses equally. ■



Lawyers for the University of California, Berkeley, and the Broad Institute faced off in patent court.

INTELLECTUAL PROPERTY

CRISPR patent battle goes to court

Hearing focuses on use of gene editing in complex cells.

BY SARA REARDON, ALEXANDRIA, VIRGINIA

It was a tough day in US patent court for the University of California, Berkeley.

On 6 December, lawyers for the university laid out its claim to the gene-editing tool CRISPR–Cas9 during a hearing at the US Patent and Trademark Office (USPTO) — and drew intense, sometimes sceptical, questioning from the three judges who will decide the fate of patents that could be worth billions of dollars.

Berkeley and its rival, the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, are each vying for the intellectual property underlying CRISPR–Cas9, which is adapted from a system that bacteria use to fend off viruses. During the hearing in Alexandria, Virginia, the USPTO judges challenged Berkeley’s central claim: that once its researchers demonstrated that CRISPR–Cas9 could be used to edit DNA in bacteria, any reasonably skilled person could have adapted the technique for use in more complex cells.

If the court decides that is true, it would invalidate the patent now held by the Broad

Institute. But the Berkeley argument is a difficult one to make, given that it hinges on “a really subjective standard” — especially when applied to extraordinarily accomplished scientists such as those at the Broad, says Jacob Sherkow, a legal scholar at New York Law School in New York City.

BYZANTINE BATTLE

The patent fight began in May 2012, when Jennifer Doudna, a molecular biologist at Berkeley, filed for a patent after her research team used CRISPR–Cas9 to alter specific stretches of bacterial DNA. In December 2012, synthetic biologist Feng Zhang of the Broad Institute filed his own patent claim, demonstrating use of the gene-editing technique in more-complex eukaryotic cells, such as those from mice and humans. Zhang asked for — and was granted — an expedited review for his patent application.

The USPTO awarded him the rights to CRISPR–Cas9 in 2014. Berkeley then asked the patent office to investigate who first invented the gene-editing technique — a process known as a ‘patent interference’. That review began in January. Over the past 11 months, the rival research institutions have filed hundreds of pages of documents with the court.

“My impression is both will end up with something.”

The 6 December hearing was the first and only time that the two sides will speak to the judges before the court rules on the patent rights. An hour before the hearing began, the line of people waiting to watch the arguments wrapped around the Christmas tree in the lobby of the USPTO and filled two overflow rooms. Each side's lawyer had only 20 minutes to present his case to the three judges.

During the hearing, the Broad's lawyer quoted liberally from news articles and interviews in which Doudna said that her lab had struggled to adapt CRISPR–Cas9 to eukaryotic cells. “This is the antithesis of something that would have been obvious,” said the Broad's lawyer, Steven Trybus.

Berkeley's lawyer Todd Walters downplayed these difficulties, saying that Doudna did not immediately publish CRISPR–Cas9 to edit eukaryotic cells because she knew it would work. Once the technology's ability to edit DNA had been proven, he told the judges, “the only thing left was to do it”.

A QUESTION OF INTENT

But the judges seemed to disagree, and grilled Walters far harder than they did Trybus, who represented the Broad. “I'm not buying that everyone who does an experiment believes it would work,” said Judge Richard Schafer. Rather, he added, a scientist such as Doudna may simply hope that her research will succeed.

This exchange suggests that Berkeley will have a hard time convincing the court that Doudna expected CRISPR–Cas9 to work in eukaryotes, Sherkow says. The university's lawyers “were trying to clarify what a biologist in 2012 would have contemplated”, he notes.

But biochemist Dana Carroll of the University of Utah in Salt Lake City, who wrote a declaration to the court on Berkeley's behalf, disagrees. “To embark on a project takes a certain amount of time, effort and money,” he says. “I don't think you'd do that unless you had some expectation of success.” He points out that several other groups began working on CRISPR–Cas9 in eukaryotes at the same time as Zhang did.

Several experts who watched the proceedings say that the Broad's prospects look brighter now, given the judges' heavy questioning of Berkeley's lawyer. “My impression is both will end up with something,” says legal scholar Robert Cook-Deegan of Arizona State University's campus in Washington DC.

The Broad has hedged its bets by filing 13 patents related to CRISPR. Several of these deal with an alternative CRISPR system in which the DNA-cutting enzyme is taken from a different species of bacteria. Because it was developed independently, Sherkow doubts that Berkeley could claim any rights to it.

He expects that the USPTO will decide the case in the next two months, although there is no deadline by which it must do so. ■

POLICY

Top US science job still in question

President-elect Donald Trump has given no clues as to whether he will appoint a science adviser.



HULTON-DEUTSCH COLLECTION/CORBIS/GETTY

Electrical engineer Vannevar Bush became the first US presidential science adviser in the 1940s.

BY ALEXANDRA WITZE

US president-elect Donald Trump has chosen people for key jobs overseeing national security, defence and environmental policy. But he has not addressed whether he will fill the most important job in US science: presidential science adviser.

Historically, many incoming presidents — who are elected in November — have designated a science adviser in December, as they move to the White House. But Trump's transition team has not contacted the White House Office of Science and Technology Policy (OSTP), which the science adviser leads, to discuss the changeover. Many researchers worry that if Trump does not pick an adviser soon, science will have a much weaker voice during the next four years.

“I have some questions as to whether Trump is going to want a science adviser at all,” says Albert Teich, a science-policy expert at George Washington University in Washington DC. “He doesn't like briefings, he doesn't like to listen to people. I can't imagine that whoever he appoints would

have a very influential position.”

Still, some of Trump's earliest moves as president may involve scientific topics. He has said that on his first day in office, 20 January, he will repeal many of the executive orders that Barack Obama has used to set policy — including those on energy and climate.

Getting a science adviser in place early would help Trump to understand the scientific implications of such issues, says

“I can't imagine that whoever he appoints would have a very influential position.”

Neal Lane, a physicist at Rice University in Houston, Texas, who advised President Bill Clinton from 1998 to 2001. “The president could make

really good use of advice from someone he has chosen who's knowledgeable about science and technology,” Lane says.

Given Trump's lack of ties to the academic or scientific communities, some speculate that he will seek technical advice from business or high-tech leaders. His transition team includes Silicon Valley billionaire Peter Thiel, who — among other things ▶

► — funds a fellowship for young adults to bypass college and develop business ventures. “We’re going to have a whole new set of people in Washington,” says Deborah Stine, a science-policy expert at Carnegie Mellon University in Pittsburgh, Pennsylvania, who served in the Obama White House for three years.

Trump may also prove open to arguments about how research can strengthen US competitiveness. Stine points to an influential report released in 2005, during George W. Bush’s administration, that described the importance of research to the national economy. Put together by a committee led by aerospace chief executive Norman Augustine, the analysis helped shape bipartisan legislation to support innovation — with strong backing from the White House.

Being named early in a president’s administration increases the chance that a science adviser can influence who will lead science agencies, and other key decisions. Presidents Clinton and Obama both chose their advisers the month after they were elected. But George W. Bush took seven months to pick physicist John Marburger. (Every presidential science adviser has been male, and most have been physicists.) By the time Marburger started the job, the Bush administration had made several crucial science-related announcements, such as restricting funding for research with human embryonic stem cells.

Many scientists criticized Marburger for serving in what some called an anti-science administration. But the adviser’s job is to provide technical input into policy decisions, not to make them, says Roger Pielke Jr, a science-policy expert at the University of Colorado Boulder. “The science adviser is not a philosopher-king,” he says.

Although the OSTP is codified in law, the president does not have to make use of it. Several members of Trump’s transition team came from the Heritage Foundation, a conservative think tank in Washington DC that issued a policy paper in June suggesting that the office be eliminated to reduce bureaucracy.

Only Congress could shrink or eliminate the OSTP. Doing so would hurt US science, says Rosina Bierbaum, an environmental scientist who headed the office for eight months in 2001 until Marburger took over. That’s because it coordinates funding for science across government agencies, and is the main entity looking for redundancies and gaps in those portfolios.

Wherever it comes from, science advice in the Trump administration will be crucial, says Lewis Branscomb, a physicist who has served in various presidential advisory groups stretching back to 1964. “The new president is going to need all the help he can get — that he will take.” ■

DRUG DEVELOPMENT

Programs face off in cancer contest

Predictive algorithms may help to whittle down the possible candidates for personalized cancer vaccines.

BY HEIDI LEDFORD

Could predictive algorithms be the key to creating a successful cancer vaccine? Two US nonprofit organizations plan to find out by pitting a range of computer programs against each other to see which can best predict a candidate for a personalized vaccine from a patient’s tumour DNA.

The Parker Institute for Cancer Immunotherapy in San Francisco, California, and the Cancer Research Institute of New York City announced the algorithmic battle on 1 December. It is part of a multimillion-dollar joint project to solve a major puzzle in the nascent field of cancer immunotherapy: which of a patient’s sometimes hundreds of cancer mutations could serve as a call-to-arms for their immune system to attack their tumours.

If the effort succeeds, it could spur the development of personalized cancer vaccines that use fragments of these mutated proteins to fire up the body’s natural immune responses to them. Because these mutations are found in cancer cells and not healthy ones, the hope is that this would provide a non-toxic way to battle tumours.

The idea is gaining traction. In 2014, news that vaccines containing such mutated proteins had vanquished tumours in mice set off a mad dash to find out whether the approach would work in people. A generation of biotechnology companies has been founded around the concept, and clinical trials run by academic labs are under way.

Still, a challenge remains. To be a good candidate for a vaccine, a mutated cancer protein must be visible to T cells, the soldiers of

the immune system. And for that to happen, tumour cells must chew up the protein into fragments. Those fragments then must bind to specialized proteins, which are shipped to the cell’s surface to be displayed to passing T cells.

The trick that vaccine researchers must master is using a tumour’s DNA to predict which mutations to home in on. “We can do the sequencing and find out the mutations, but it’s very hard to know which of these tens or hundreds or thousands of mutations are actually going to protect people from the growth of their cancers,” says Pramod Srivastava, an immunologist at the University of Connecticut School of Medicine in Farmington.

One approach is to use algorithms to predict which bits of a mutated protein might be seen by a T cell.

“It’s very hard to know which of these tens or hundreds or thousands of mutations are actually going to protect people.”

These work by analysing where the proteins could be cleaved, for example, and which of the resulting fragments will bind tightly to the molecules that put them on display.

But each laboratory has a different “secret sauce”, says Robert Schreiber, a cancer immunologist at Washington University in St. Louis, Missouri. And most are not very predictive: Robert Petit, chief scientific officer of biotechnology company Advaxis in Princeton, New Jersey, estimates that the algorithms are typically less than 40% accurate.

To solve the problem, the Parker Institute and the Cancer Research Institute launched their challenge. They have arranged for



MORE ONLINE

STAY CURRENT

- Graphene-spiked Silly Putty picks up human pulse go.nature.com/2hscpvv
- LIGO echoes hint at relativity breakdown go.nature.com/2hhhhxr
- Fingernail rules out lead-poisoning death of Arctic explorer go.nature.com/2hzpesf

NATURE PODCAST

A spray that boosts plant growth and resilience; 3-million-year old hominin footprints; and the seahorse genome nature.com/nature/podcast



30 laboratories that already use such algorithms to apply their secret sauces to the same DNA and RNA sequences. The sequences will come from cancers such as melanoma and lung cancer, which tend to have many hundreds of mutations (see ‘Mutation map’) and thus could provide ample possibilities for a vaccine.

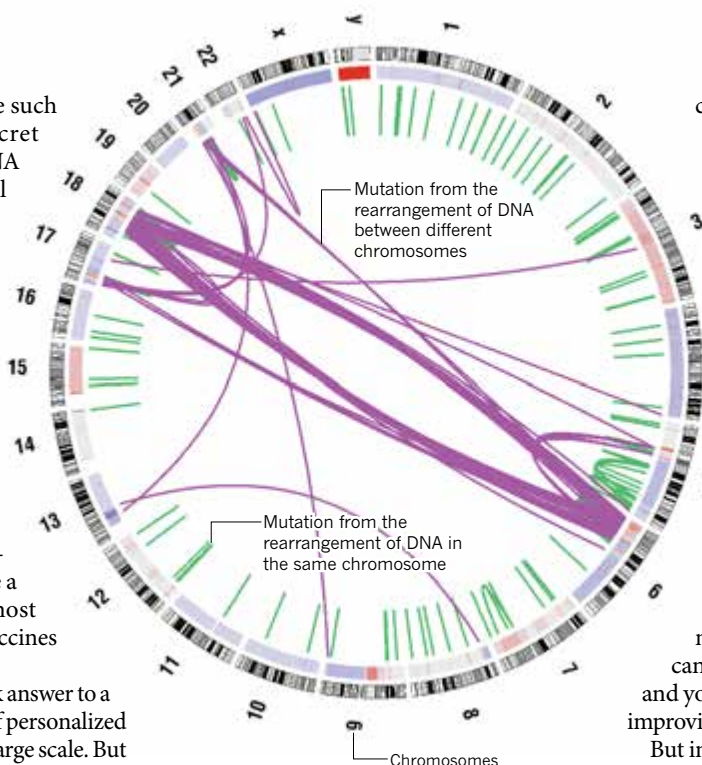
A handful of other laboratories will then test whether any T cells in the tumour recognize those fragments, and are stimulated by them — a sign of a good vaccine target. The alliance will not publicly announce a winner, but hopes to use the most accurate algorithms to design vaccines for clinical trials.

Algorithms can provide a quick answer to a complicated question — crucial if personalized vaccines are to be deployed on a large scale. But ultimately, Srivastava says that the best way to improve the algorithms is to collect more data from animal studies to learn about how T cells naturally respond to mutations. His lab and others are making hundreds of putative vaccines tailored to an individual tumour, and administering them to mice to see which are

capable of fighting the cancer.

And Drew Pardoll, a cancer immunologist at Johns Hopkins University in Baltimore, Maryland, worries that algorithms may never account for some factors that influence T-cell responses. For example, mutations may be less suitable for a vaccine if they have arisen early in tumour development, giving the immune system time to begin viewing them as ‘normal’. Pardoll argues that the field needs faster, easier and more accurate laboratory tests to determine which mutations best trigger a T-cell response. “We don’t yet know enough about the rules to make perfect predictions,” he says. “You can algorithm until the cows come home and you’re not really going to know if you’re improving things.”

But in the absence of speedy lab tests, companies need algorithms, argues Robert Ang, chief business officer at Neon Therapeutics of Cambridge, Massachusetts. “There is already evidence to show that this approach works despite the imperfect algorithms,” he says. “Improving the algorithms even more could be very meaningful.” ■



MUTATION MAP

The hundreds of mutations in the genome of a melanoma tumour could be used to induce an immune attack on cancer cells without harming healthy cells — and can be visualized as a map.

The myth buster

*Hans Rosling is on
a mission to save
the world from
preconceived ideas.*

BY AMY MAXMEN



Hans Rosling knew never to flee from men wielding machetes. “The risk is higher if you run than if you face them,” he says. So, in 1989, when an angry mob confronted him at the field laboratory he had set up in what is now the Democratic Republic of the Congo, Rosling tried to appear calm. “I thought, ‘I need to use the resources I have, and I am good at talking.’”

Rosling, a physician and epidemiologist, pulled from his knapsack a handful of photographs of people from different parts of Africa who had been crippled by konzo, an incurable disease that was affecting many in this community, too. Through an interpreter, he explained that he believed he knew the cause, and he wanted to test local people’s blood to be sure. A few minutes into his demonstration, an old woman stepped forward and addressed the crowd in support of the research. After the more aggressive members of the mob stopped waving their machetes, she rolled up her sleeve. Most followed her lead. “You can do anything as long as you talk with people and listen to people and talk with the intelligentsia of the community,” says Rosling.

He is still trying to arm influential people with facts. He has become a trusted counsellor and speaker of plain truth to United Nations leaders, billionaire executives such as Facebook’s Mark Zuckerberg and politicians including Al Gore. Even Fidel Castro called on the slim, bespectacled Swede for advice. Rosling’s video lectures on global health and economics have elevated him to viral celebrity status, and he has been listed among the 100 most influential people in the world by the magazines *Time* and *Foreign Policy*. Melinda Gates of the Bill & Melinda Gates Foundation says, “To have Hans Rosling as a teacher is one of the biggest honours in the world.”

But among his fellow scientists, Rosling is less popular. His accolades do not include conventional academic milestones, such as massive grants or a stream of publications in top-tier journals. And rather than generating data, Rosling has spent the past two decades communicating data gathered by others. He relays facts that he thinks many academics have been too slow to appreciate and argues that researchers are ignorant about the state of health and wealth around the world. That’s dangerous. “Campuses are full of siloed people who do advocacy about things they don’t understand,” he says.

So now, in the sunset of his career, Rosling is writing a book with his son Ola and his daughter-in-law Anna Rosling Rönnlund to dispel outdated beliefs. It has the working title *Factfulness*, and they hope it will inform everyone from schoolchildren to esteemed experts about how the world has changed: how the number of births per woman worldwide has dropped over the past few decades, for example, and how average life expectancy (71 years) is now closer to that of the country with the

highest (Japan, 84) than the lowest (Swaziland, 49). He reasons that experts cannot solve major challenges if they do not operate on facts. “But first you need to erase preconceived ideas,” he says, “and that is the difficult thing.”

LIFE ON THE BRINK

Rosling’s ambitions were born from curiosity. As a young boy in Uppsala, he listened intently as his father, a coffee-factory employee, described the hardships of the East African labourers who picked the beans. Rosling and his girlfriend, Agneta Thordeman, joined student protests against South African apartheid and the US war in Vietnam.

The couple studied medicine — she as a nurse and he as a doctor — and travelled through India and southeast Asia on a shoestring budget. In 1972, they were married and seven years later they moved to Mozambique with their two small children.

“Extreme poverty produces diseases. Evil forces hide there.”

Rosling wanted to fulfil a promise he had made many years earlier to the founder of the Mozambican Liberation Front, Eduardo Mondlane. Mondlane had explained that Mozambique’s future would be challenging after the country gained independence from Portugal, because the nation was so poor and education levels low. Rosling recalls, “He shook my hand and looked me in the eyes and said: ‘Promise you will work with us.’” Mondlane was killed by a letter bomb soon afterwards — he did not live to see independence, which came in 1975 — but Rosling kept his word.

The Mozambican government assigned Rosling to a northern part of the country, where he would be the only doctor serving 300,000 people. Because of the scarcity of health care, patients were often in excruciating pain by the time he saw them. Rosling recalls performing emergency surgery to extract dead fetuses from women on the verge of death. He watched helplessly as children perished from diseases that should have been simple to prevent. “Those years became a sort of trauma,” he says.

In 1981, he received a letter from an Italian nun working as a nurse at a remote health post. “Please come,” she wrote. People in the surrounding villages had been stricken with sudden paralysis of both legs. Separating from his family, Rosling embedded himself in the crisis.

He was assigned to lead a survey of

500,000 people and found that populations with the highest rate of the disease survived entirely on bitter cassava, the only crop that could grow when drought struck the region. The plant turned out to contain cyanogenic glucoside, a precursor to cyanide. Typically, soaking cassava roots in water for several days removed the toxin. But with streams running dry and families starving, women who prepared cassava had skipped this step — to their detriment. Dietary amino acids can also detoxify the poison, but people had no access to meat or beans that provide them.

At the end of 1981, owing to a number of circumstances including the death of their third child, Rosling and his family returned to Sweden. Rosling became a lecturer on health care in low-income countries at Uppsala University but spent time in Tanzania and the Congo region as well, studying the paralysing disease he had first observed in Mozambique. He noticed that no matter what country he was in, the towns afflicted looked similarly tragic. Skeleton-thin people hobbled down dirt paths on makeshift crutches, or crawled with their legs twisted and dangling behind them like anchors. One Congolese community called the malady konzo, derived from a word referring to an antelope tethered at its knees. This is the name that Rosling would use in 1990, when he and his colleagues formally defined the disease and laid out the evidence for what causes it (W. P. Howlett *et al. Brain* **113**, 223–235; 1990).

As Rosling travelled, he trained African graduate students who specialized in konzo, and together they found that proper cassava processing was the most realistic method of short-term prevention. However, the message often fell on deaf ears because of hunger and conflict. Rosling became convinced that the real root of konzo resided not in cassava, but in economic devastation. “Extreme poverty produces diseases. Evil forces hide there,” he says. “It is where Ebola starts. It’s where Boko Haram hides girls. It’s where konzo occurs.”

THE TRUE PICTURE OF POVERTY

The World Bank defines extreme poverty as a state in which people survive on less than US\$1.90 per day. Rosling can recognize it in other ways. He has seen it in people who must walk for hours without shoes to find water or to farm eroded soil. He sees it in those who remain short because of malnourishment, whose babies are born dangerously underweight and who are trapped with no options in life.

Ultimately, he says that eliminating extreme poverty is the only way to cure konzo and prevent other maladies — both social and infectious. Money, politics and culture underlie disease in many circumstances, he argues.

Take an outbreak in Cuba that Rosling investigated in 1992. The Cuban embassy in Sweden had asked him to find out whether toxic cassava could have caused roughly 40,000 people



Rosling is known for his creative use of visual aids, from sophisticated animations to children's toys.

to experience visual blurring and severe numbness in their legs. On his first morning in Havana, Rosling met local epidemiologists in a conference room. "Then, two men walk in with guns, and in comes Fidel Castro," he recalls. "My first surprise was that he was so kind, like Father Christmas. He didn't have the attitude you might expect from a dictator."

With Castro's approval, Rosling travelled to the heart of the outbreak, in the western province of Pinar del Río. It turned out that there was no link with cassava. Rather, adults stricken with the disorder all suffered from protein deficiency. The government was rationing meat, and adults had sacrificed their portion to nourish children, pregnant women and the elderly.

Reporting back to Castro, Rosling couched his conclusions carefully: "I know your neighbours want to force their economic system on you, which I don't like, but the system needs to change because this planned economy has brought this disease to people." After his presentation, Rosling went to the toilet. A Cuban epidemiologist approached him to thank him. He and his colleagues had come to the same conclusion several months earlier, but they were removed from the investigation for criticizing communism. Corroboration of their work from Rosling and other independent researchers supported the policy changes that stemmed the outbreak.

IGNORANCE ABOUT IGNORANCE

Back in Sweden, Rosling continued to teach global health, moving to the Karolinska Institute in Stockholm in 1996. But he came to realize that neither his students nor his colleagues grasped extreme poverty. They pictured the poor as almost everyone in the 'developing world': an arbitrarily defined territory that

includes nations as economically diverse as Sierra Leone, Argentina, China and Afghanistan. They thought it was all large family sizes and low life expectancies: only the poorest and most conflict-ridden countries served as their reference point. "They just make it about us and them; the West and the rest," Rosling says. How could anyone hope to solve problems if they didn't understand the different challenges faced, for example, by Congolese subsistence farmers far from paved roads and Brazilian street vendors in urban *favelas*? "Scientists want to do good, but the problem is that they don't understand the world," Rosling says.

"Global health seems to have entered into a post-fact era."

Ola, his son, offered to help explain the world with graphics, and built his father software that animated data compiled by the UN and the World Bank. Visual aids in hand, the elder Rosling began to script the provocative presentations that have made him famous. In one, a graph shows the distribution of incomes in 1975 — a camel's back, with rich countries and poor countries forming two humps. Then he presses 'go' and China, India, Latin America and the Middle East drift forward over time. Africa moves ahead too, but not nearly as much as the others. Rosling says, "The camel dies and we have a dromedary world with one hump only!" He adds, "The per cent in poverty

has decreased — still it's appalling that so many remain in extreme poverty."

Rosling's online presentations grew popular, and the investment bank Goldman Sachs invited him to speak at client events. His message seemed to support advice from the firm's chief economist, Jim O'Neill. In 2001, O'Neill had coined the acronym BRIC for the emerging economies of Brazil, Russia, India and China, often considered part of the developing world. He warned that financial experts ignored these rising powers at their peril. "I used to tease my colleagues who thought in a traditional framework," O'Neill says. "Why are we talking about China as the developing world? Based on the rate of economic growth, China creates another Greece every three months; another UK every two years."

Rosling welcomed the new audience. "They request my lectures because they want to know the world as it is," he says. The private sector needs to understand the economic and political conditions of current and potential markets. "To me it was horrific to realize that business leaders had a more fact-based world view than activists and university professors."

O'Neill left Goldman Sachs in 2013, and went on to lead a committee on global antibiotic resistance. He looked to Rosling for a big-picture view. "I wish there were more people like him," says O'Neill. "He genuinely thinks about the future of all seven-plus-billion of us, rather than so many who claim they do but actually come at it with a narrow and national perspective."

Rising wealth pleases Rosling because he wants extreme poverty to disappear. To help get there, he celebrates improvements. He calls the UN's push to eradicate extreme poverty by 2030 an entirely reasonable goal because the proportion of people living in extreme poverty has declined by more than half in the past quarter of a century, and the strategies needed to help the remainder are known.

His attitude aligns him with Steven Pinker of Harvard University in Cambridge, Massachusetts, who wrote *The Better Angels of our Nature* (Viking, 2011). In the book, Pinker argues that global rates of violence are much lower than they were in the past. The two met at a TED conference in 2007, when Pinker took the stage after Rosling ended his talk by swallowing a sword (whatever grabs attention). Pinker says that Rosling made him think that "the decline in violence might be a part of an even bigger story about humans gradually making progress against other scourges of the human condition".

Both have been criticized as being Pollyannaish about the global situation in the face of tragedies such as the conflict in Syria. "People think that if you emphasize how things have gone well it is the same as saying no problems remain. That's not true," Pinker counters. "In fact, I strongly suspect that people are more motivated to reduce problems like poverty and

violence if they think there is a good chance they can succeed.”

And as a cognitive scientist, Pinker admires the animations that Rosling uses. One, which depicts countries as bubbles that migrate over time according to wealth, life span or family size, allows viewers to grasp multiple variables simultaneously. “It’s a stroke of genius,” Pinker says. “He gets our puny human brain to appreciate five dimensions.”

In 2005, Rosling, Ola and Anna founded the non-profit Gapminder Foundation in Stockholm to develop the ‘moving-bubble’ software, Trendalyzer, and to spread access to information and animated graphs depicting world trends. Google acquired Trendalyzer in 2007, and Gapminder has successfully pressured the World Bank to make its data free to the public.

HOW TO DISMANTLE THE POPULATION BOMB

Rosling’s charm appeals to those frustrated by the persistence of myths about the world. Looming large is an idea popularized by Paul Ehrlich, an entomologist at Stanford University in California, who warned in 1968 that the world was heading towards mass starvation owing to overpopulation. Melinda Gates says that after a drink or two, people often tell her that they think the Gates Foundation may be contributing to overpopulation and environmental collapse by saving children’s lives with interventions such as vaccines. She is thrilled when Rosling smoothly uses data to show how the reverse is true: as rates of child survival have increased over time, family size has shrunk. She has joined him as a speaker at several high-level events. “I’ve watched people have this ‘aha’ moment when Hans speaks,” she says. “He breaks these myths in such a gentle way. I adore him.”

The appreciation extends to the World Health Organization: director-general Margaret Chan says that Rosling provides facts for decision-makers to consider. “He makes the case that as people grow in wealth, they grow in health,” she says. And his talks help her to convince governments that data collection can help them to track whether they are getting returns on their investments in global health.

The past few years have brought new challenges. In 2014, Ebola was spreading in West Africa, and Rosling’s liver was failing. A hepatitis C infection that he had mysteriously acquired in his youth was becoming lethal. He travelled to Japan to receive the newest treatment, not yet approved in Sweden. By October, he found himself fretting, from afar, over discrepancies in official reports on the number of suspected and confirmed Ebola cases. “I realized my skills were needed,” he says.

As soon as the drugs cured him, Rosling flew to West Africa to join the Liberian government’s epidemiological-surveillance team. The team wanted to consolidate data, but struggled with the disparate ways in which international agencies collected information. “We

QUIZ

Test your world knowledge

In some of his talks, Hans Rosling likes to explore the audience members’ misconceptions about the world. He finds that people often perform worse than predicted by chance.

In the past 20 years, the proportion of the world population living in extreme poverty has roughly...

- Doubled
- Remained the same
- Decreased by 10%
- Decreased by half

Globally, men aged 25 and older have spent about 8 years in school on average. How many years have women that age spent in school?

- 2 years
- 3 years
- 5 years
- 7 years

➔ **NATURE.COM**

For answers and more quiz questions, visit:
go.nature.com/2gkhqxl

were losing ourselves in details,” says Rosling. “I saw this was a war situation: all we needed to know is, are the number of cases rising, falling or levelling off?” After a few months, it became clear that the rate of new cases had diminished. Rosling was rewarded with a traditional chieftainship by the Liberian government.

Now, at the age of 68, Rosling has retreated to his red wooden house in Uppsala with Agneta. He continues to work and plugs away at his “factfulness book on megamiskonceptions”. Every now and again, he stirs the pot. In October, he published a piece in *The Lancet* identifying a misleading statistic in a widely cited report from an advocacy organization launched by the UN (H. Nordenstedt and H. Rosling *Lancet* 388, 1864–1865; 2016). The group claimed that 60% of maternal deaths occur in settings of conflict, displacement and natural disaster. Rosling checked the numbers and calculated that the true amount was no more than 17%. A UN spokesperson explains that part of the discrepancy derives from the fact that in the original figure, women who gave birth in nations affected by crises were included — even if their region had not been directly impacted.

Rosling blames the popularity of the dramatic-sounding statistic on the desire to raise funds at a time when refugee crises garner financial support. “Global health seems to have entered into a post-fact era, where the labelling of numerators is incorrectly tweaked for advocacy purposes,” he wrote in the *Lancet* article with Helena Nordenstedt, a colleague at the Karolinska Institute. The majority of maternal deaths occur among the extremely poor, they added. Those remote populations are hidden even from the aid community.

Rosling prods academics when he can (see ‘Test your world knowledge’). For instance, at a Nobel-laureate meeting in Lindau, Germany, in 2014, he quizzed the audience of leading scientists on the average life expectancy in the world today. Out of three choices, just over one-quarter of the crowd picked the correct answer of 70. That’s less than would be expected by chance. The quiz spurred laughter in Lindau, but scientists are generally not his audience. Rosling is rarely invited to give keynote lectures or departmental seminars because he doesn’t push a single field forward; he has not made fundamental discoveries since his konzo days. Researchers agree that he is a good communicator — but not the kind to teach scientists.

“People like Hans Rosling face the criticism of being too superficial,” explains Peter Hotez, a tropical-disease scientist at Baylor College of Medicine in Houston, Texas. “It’s the dilemma of the public intellectual,” he says, describing academics who bridge several disciplines rather than excel at one.

Rosling says he never cared much about his academic reputation. He was lucky to receive steady support from the former head of the Karolinska Institute, Hans Wigzell, who encouraged him to seek outside funding so that he could pursue whatever he deemed most important. After Rosling decided that that meant teaching broadly, he walked away from research entirely.

He also differs from global-health experts who have stepped outside academia to change policies. He hasn’t worked to expand access to HIV medication, for example. He has not — like Hotez — put neglected tropical diseases on the world health agenda. And konzo still exists. But Rosling has had success; it’s just that the impact becomes harder to measure the broader his goals become. Now that he has decided that the public at large must buy into ending extreme poverty and creating a sustainable world, he has dedicated the last chapter of his career to education. With the right facts, he hopes, people will make the right decisions — he just needs to face down the misconceptions.

Who is better suited to the task than a man able to stave off machetes with the power of a few pictures and his words? ■

Amy Maxmen is a science journalist in Berkeley, California.

COMMENT



HISTORY The role of scurvy in the age of discovery **p.338**

SUSTAINABILITY The state of the Earth looks rosy, on a scale of billions of years **p.339**

EDUCATION How is the digital revolution working out for students and lecturers? **p.340**

OBITUARY Ralph Ciceone, environment-defending NAS head, remembered **p.342**

DIETER TELEMANS/PANOS



A cataract patient receives synthetic lenses in the Democratic Republic of the Congo.

Make better, safer biomaterials

Design and test new polymers with clinical uses in mind, urge **Nicholas A. Peppas** and **Ali Khademhosseini**.

Polymers have a wide range of physical and mechanical properties suited to many purposes in medicine. For example, poly(methyl methacrylate) (PMMA), which resembles bones and teeth, has been used since the 1930s for dental implants and hip replacements. Poly(2-hydroxyethyl methacrylate) has been used since the early 1960s for soft contact lenses because it is transparent, flexible and stays swollen and wet. Strong yet bendable polyurethanes have been used for heart valves for decades.

But once in the body, polymers can cause side effects. These might be triggered by components left over from the polymerization process, such as monomers, reaction initiators or catalysts. For example, residual methyl methacrylate monomer in PMMA damages cells, irritates eyes and skin and disturbs the nervous system¹. Certain silicones in breast implants can cause infections². Poly(ethylene terephthalate), often used to make vascular grafts, traps proteins on its surface that can disturb blood flow and induce clots³.

Clinical approval of new materials remains difficult and expensive⁴. Rounds of extensive toxicological studies are followed by tough clinical trials to assess the safety and efficacy of a proposed device. These hurdles mean that repurposing old materials is easier than introducing new ones. But promising new options abound.

What's needed is a more integrated approach to designing and regulating polymers in biomedicine. From the start, designers need to address all the components ▶

► that may render a material toxic or capable of causing cancer, birth defects, genetic mutations or blood clots. Below we outline the sort of standardized testing platforms — experimental and computational — that are needed to evaluate biocompatibility.

ADVANCED POLYMERS

Progress over the past two decades holds promise for designing new biomaterials. For example, macromolecular structures can be designed and fabricated with precision. Techniques such as reversible-deactivation radical polymerization attach and detach small active molecular groups (radicals) to block undesirable steps during the reaction that forms the polymer. The range of molecular weights of the polymer chains is controlled and little catalyst is left behind (just a few parts per million). The biocompatibility of the polymers can be enhanced by further purification and by using aqueous solvents and non-metal catalysts.

Another breakthrough is ‘click chemistry’, which builds polymers and molecules in a modular way through a series of reactions. Different sorts of polymer can be linked together, vastly broadening the range of surface chemistries possible for biomedicine. Click reactions are efficient, have high yields and few by-products⁵. They need only mild conditions and benign or removable solvents. Click chemistry has been used to make gels, including patterned forms of hydrogel, where different areas perform different biochemical functions⁵. Unfavourable copper catalysts and azides are being phased out through, for example, carrying out reactions with greater precision and without catalysts.

Another emerging area is assembly through physical interactions between molecules. For example, hydrophilic and -phobic groups arrange themselves differently in polar or non-polar liquids. They can self-assemble into thin plates, aggregates and three-dimensional structures. Biopolymers made from DNA and proteins are increasingly used to make materials. Polymeric materials shaped as nanotubes, nanospheres, fibres and tapes have been prepared by self-assembling peptides or macromolecules.

Surfaces can be modified to control interactions. For example, some hydrogels repel proteins electrostatically, which avoids immune reactions or the biosurface becoming fouled⁶. Repellent coatings on medical devices such as catheters can be based on slippery, liquid-infused, porous surfaces (SLIPS) to prevent thrombosis⁷.

“It is easier to use established polymers in new applications than to get new ones approved”



Silicone membranes for breast implants are tested for water resistance.

APPROVAL PROCESS

However, it is difficult to get clinical approval for new polymeric systems⁴. Toxicological studies and clinical trials require more money and equipment than a standard academic laboratory can muster. An industrial setting is a must.

The regulatory process for multifunctional medical devices is complicated. For example, a single product such as a heart stent that slowly releases a drug can have several components, including the drug, a polymer coating and metallic frame. The US Food and Drug Administration evaluates the effectiveness and safety of either the primary use or of independent uses, depending on what the product is mainly meant to do. Thus it is easier to use established polymers in new applications than to get new ones approved.

Addressing these challenges requires various stakeholders to work together, including academia, industry and regulatory agencies. They should evaluate the biomaterial's design earlier in the research phase, based on regulatory needs and the performance specifications required for each application.

EVALUATION SYSTEMS

To predict how human tissues will respond to new materials before they are tested in clinical trials we must develop standardized *in vitro* and *in vivo* evaluation platforms. Toxicity and inflammatory response are particularly important to assess because these human reactions cannot be faithfully reproduced in animal models. Several options need research.

‘Organ-on-a-chip’ systems capture aspects of human physiology using miniaturized human tissues. Most existing platforms focus on metabolic and barrier tissues, such as liver and lung. Polymeric materials will

require systems capable of testing human blood and the immune system. Variations in patients must be considered. Organ-on-a-chip could contain cells derived from specific patients.

Modelling and simulations can be used to understand and predict the behaviour of body systems, from a single cell to the whole organism. These computational tools can also be used to interpret, analyse and predict the underlying response mechanisms to new substances. For example, a mathematical method known as physiologically based pharmacokinetic modelling (PBPK) is used to determine exposure doses that can lead to toxicity⁸. PBPK parameters use data from studies done *in vitro*, *in vivo* and *in silico*. The increased risk of human cancer from vinyl chloride was evaluated using such a model. Although based on data from animals, its estimates are consistent with human epidemiological data.

High-throughput screening can be used to test the safety of libraries of new polymers, while lowering the cost and reducing animal testing. To screen hundreds of thousands of chemicals, rapid throughput microarrays have been used *in vitro* and *in vivo*⁹. Automatic systems that contain small vertebrates to conduct high-throughput pharmacological tests *in vivo* have been developed¹⁰. Similar tools would be useful in assessing new polymers.

Integrating all these platforms into the design process would accelerate the clinical translation of biomaterials. ■ [SEE INSIGHT P.352](#)

Nicholas A. Peppas is professor in the Departments of Biomedical Engineering, Chemical Engineering, Pharmacy and Surgery and Perioperative Care at the Dell Medical School of The University of Texas at Austin, USA. **Ali Khademhosseini** is professor in the Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. e-mail: peppas@che.utexas.edu

- Gupta, S. K., Saxena, P., Pant, V. A. & Pant, A. B. *Toxicol. Int.* **19**, 225–234 (2012).
- Pittet, B., Montandon, D. & Pittet, D. *Lancet Infect. Dis.* **5**, 94–106 (2005).
- Langer, R. & Tirrell, D. A. *Nature* **428**, 487–492 (2004).
- Pashuck, E. T. & Stevens, M. M. *Science Transl. Med.* **4**, 160sr164 (2012).
- DeForest, C. A., Polizzotti, B. D. & Anseth, K. S. *Nature Mater.* **8**, 659–664 (2009).
- Zhang, L. *et al. Nature Biotechnol.* **31**, 553–556 (2013).
- Leslie, D. C. *et al. Nature Biotechnol.* **32**, 1134–1140 (2014).
- Clewell, H. J., Gentry, P. R., Gearhart, J. M., Allen, B. C. & Andersen, M. E. *Sci. Total Environ.* **274**, 37–66 (2001).
- Kenny, H. A. *et al. Nature Commun.* **6**, 6220 (2015).
- Pardo-Martin, C. *et al. Nature Meth.* **7**, 634–636 (2010).

Mariners' malady

Tilli Tansey on how the ruinous trajectory of scurvy marked the age of discovery.

It was called the 'scourge of the sea'. Scurvy — a painful, weakening disease in which limbs may swell and gums rot — became prevalent among sailors in the early eighteenth century as adventurers voyaged around the globe. By 1800, citrus fruit was widely used to cure or even prevent the disease; the active principle, vitamin C, was finally identified in the 1930s. But as cultural historian Jonathan Lamb's intriguing *Scurvy* reveals, there is vastly more to this malady.

The key is in the subtitle. Lamb reveals a wider perspective on the disease in the context of the creation of new knowledge, as a number of primarily European explorers encountered new lands such as Australia, with new peoples, flora, fauna and foodstuffs. The story of scurvy is also entwined with medical developments, including the rise in theories about what causes disease, such as contagion, and the long search for treatments.

Lamb draws widely on explorers' records, including the diaries and logbooks of captains James Cook and William Bligh, as well as the works of ships' surgeons, apothecaries and natural philosophers such as Erasmus Darwin. Writers as diverse as Herman Melville and Nancy Mitford offer striking passages. The figure of Death in Samuel Taylor Coleridge's 1798 poem *The Rime of the Ancient Mariner* clearly displays symptoms of the disease. And in George Orwell's

Nineteen Eighty-Four (1949), a starving Winston Smith weeps as his inquisitor pulls a tooth from his scurvy-destroyed gums.

The route to a cure was riddled with detours, and a lack of consensus lasted into the twentieth century. In the early 1790s, for instance, British physician Thomas Beddoes — inspired by the discovery of oxygen and

a nitric acid 'cure' for syphilis — searched for an acidic gas to combat scurvy, tuberculosis and catarrh. Although he found no remedy, his laboratory did manufacture nitrous oxide, subsequently investigated by Humphry Davy (M. Peplow *Nature* 533, 175–176; 2016). Some believed that scurvy was caused by a poison arising from rotting provisions or bad air; others, that an invisible element necessary for life was somehow lost. Captain Cook insisted on clean, warm clothing, dry, hygienic ships and plenty of rest to prevent an outbreak.

Other captains and naval surgeons, including James Lind and particularly Thomas Trotter, offered evidence that regular provision of fresh food could halt or prevent the disease, and that specific preparations involving ingredients such as citrus fruits might be curative. Verification was inconsistent, because the supply of these antiscorbutics was often irregular, or the quality poor. By the end of the eighteenth century, concentrated lemon juice was routinely issued on British naval vessels — but even then, its efficacy was neither completely proved nor accepted.

One difficulty was that the disease was viewed as a badge of dishonour, and often denied or disguised by ships' officers. Among the many challenges faced by Bligh on the infamous voyage of the *Bounty* (which ended in mutiny in 1789) was scurvy among the crew. "A disgrace to a ship," Bligh called it, claiming that the symptoms were due to rheumatism. More than a century later, in the Antarctic, Captain Robert Falcon Scott of the *Discovery* emphasized in his log that the "great thing is to pretend that there is nothing to be alarmed at". That attitude was later adopted when convenient by authorities, including the British Ministry of Health during the Second World War, when scurvy among civilians was reported. The ministry deemed the cause to be wilful self-neglect rather than chronic food shortages.

Lamb interprets 'scurvy' broadly, perhaps wisely making no attempt at retrospective differential diagnosis. Purists may baulk at his inclusion of pellagra, beriberi and other disorders of malnutrition, but



Scurvy: The Disease of Discovery
JONATHAN LAMB
Princeton University Press: 2016.

that allows for a richer range of material and interpretation. For example, the chapter on 'scurbutic nostalgia' — the psychological and emotional impacts of the disease, including hallucinations of food, water or home — is woven through an examination of the depression attributed to 'calenture', or sea-fever.

This book is not, however, an easy read. The erudition and calls to multiple authorities can be wearisome. Take, for instance, the formulation: "The importance of air ... led by way of MacBride's notion of the fixed air in malt to the pneumatic theory of Beddoes, espoused with certain qualifications by Trotter". Accounts of enforced seafarers, especially slaves and convicts, make for harrowing reading, as do the descriptions of skin eruptions and vomiting. In places, the narrative is weirdly disturbing. During John Davis's horrific sixteenth-century voyage on the *Desire* from South America, for instance, the only food was dried penguin. This rotted and became infested with maggots that devoured almost everything on the ship, including the flesh of living men and the wood of the vessel.

Not surprisingly, given the size, scope and ambition of *Scurvy*, there are irritations. One is the sometimes anachronistic use of terms, for instance in the context of Cook's reliance on malt as an antiscorbutic "even though it contained no vitamin C". Lamb also uses the term interchangeably with 'ascorbic acid' (its chemical name) or 'ascorbate', any salt of ascorbic acid. He attempts physiological and pathological explanations that 'ascorbate' is essential for a healthy nervous system because of its role in, for example, the synthesis of dopamine and 5HT, but these are weak. And an appendix by neuroscientists James May and Fiona Harrison adds little. It is neither helpful nor revealing to extrapolate from a knockout vitamin-C-deficient mouse model to eighteenth-century seafarers with several confounding factors and other illnesses. Indeed, these efforts at modern scientific explanations detract from the richness of Lamb's cultural explorations of discovery and knowledge.

Describing *Scurvy* as a 'bit of a lemon' would be glib and unfair. It is much better than that, but remains something of a curate's egg. ■

Tilli Tansey is professor of the history of modern medical sciences at Queen Mary, University of London.
e-mail: t.tansey@qmul.ac.uk



The effects of scurvy, drawn by surgeon Henry Mahon of the HMS Barrosa in 1842.



A university lecturer being filmed for a learning website.

DIGITAL EDUCATION

Pedagogy online

Mike Sharples weighs up a study on the great migration to digital education, from ‘flipped’ teaching to MOOCs.

In 1993, educational technologist Seymour Papert suggested that a teacher from the nineteenth century transported into the mid-1990s would feel at home in the classroom. Twenty years on, this is no longer true.

Teachers in much of the developed world now use smartboards, tablets and student-centred, collaborative and project-based learning. Universities are adopting flipped teaching: students learn online, then solve problems in the classroom. Some can access remote lab equipment and telescopes. Some institutions — such as the University of Waterloo in Canada and Massey University in Palmerston North, New Zealand — blend online and campus teaching. Massive open online courses (MOOCs) involve people around the world in study and conversation. The continuing change is provoking existential dread among some faculty members, who envision teachers replaced with computer-based tutors and universities moving to online-only courses in the next decade.

Those shifts can also foster an excitement that Robert Ubell’s *Going Online* captures. The book is the view from the control room of the New York University Tandon School of Engineering, where Ubell heads the digital-education unit. He starts by observing that traditional university education has failed to engage students in active learning. The more accomplished the lecture, for instance, the more it may give a false impression that all the students have absorbed the material.

Ubell’s proposition is that online learning lets students process information in their own time. They can take part in online discussions and ask questions anonymously, without losing face. This demands a new pedagogy

— teaching, learning and assessment for active learning communities. Academics must work with web designers and educational technologists to create conditions that let students control the pace and delivery of learning, yet continually share and respond to others’ ideas.

Ubell is right that anonymity can help students who are less confident, or not fluent in the language. But an important part of university is learning to challenge and debate. Some MOOC platforms, such as FutureLearn, promote constructive discussion, with thousands of learners bringing global perspectives to hotly debated topics such as climate change.

Going Online shows there are many ways to migrate education to the Internet. All require institutions to commit to opening up instruction, moving from a professional relationship between a teacher and students to a corporate process. It involves decisions about the online learning environment (be it Moodle, Blackboard or Canvas), whether to use a MOOC provider, how to negotiate intellectual-property rights and how to compensate staff. In offering students autonomy and activity, the online university may sacrifice humanity.

The way back from the mass corporate online instruction offered by some for-profit universities, such as the University of Phoenix in Arizona, is through blended learning. Students study the curriculum online from material provided by sources including MOOCs, web pages and interactive science simulations. They are encouraged to use social media to share knowledge. The classroom becomes a site for exploring a topic in depth by solving

**Going Online:
Perspectives on
Digital Learning**
ROBERT UBELL
Routledge: 2016.

problems, debating and taking tests. In science, students can get hands-on experience with lab equipment, and then book remote access and analyse data online. Blended learning works equally well for apprenticeships and professional development. The Swiss government’s DUAL-T online vocational-learning initiative, for example, bridges the gap between classroom and workplace.

An academic who has spent a career lecturing may be uncomfortable with the shift to facilitating learning, but new teachers have grown up with online learning and social media. Many will have used collaboration tools like Slack, and professional communities such as LinkedIn and Stack Exchange.

At the centre of the book is a 2000 study by Ubell and his colleague Hosein Fallah that compares two graduate classes — identical in content and instructor, but with one delivered through lectures and the other online. The numbers are small (just 7 students online and 12 on campus), and the results inconclusive. A better demonstration is a metastudy led by educational psychologist Barbara Means (mentioned briefly in the book) that analysed more than 1,000 empirical studies. It found that, on average, students engaged in online learning did better than those who had solely face-to-face instruction. The advantage was bigger for blended learning (B. E. Means *et al. Evaluation of Evidence-Based Practices in Online Learning*; US Department of Education, 2009).

As Ubell says, critics of online learning generally point to training systems and MOOCs that deliver canned lectures. Success in digital education comes from social-networked learning, with global access to online materials, high-quality open courses and vibrant peer discussions. The flipped classroom can work in both New Delhi and New York City. It requires a decentred perspective to create communities of education providers and learners, welcoming differing cultural perspectives and pedagogies. The pioneers are universities committed to global open education, such as the Open University in Milton Keynes, UK; the Massachusetts Institute of Technology in Cambridge; Canada’s Athabasca University; and the University of Cape Town in South Africa. The most traditional universities are finding this step the hardest.

Just as modern education is becoming a melange of sources and services, so *Going Online* is pieced together from previously published, updated papers. Weaving a coherent narrative can be challenging, but the book captures aspects of an education system in transition from campus instruction to global enterprise. ■

Mike Sharples is professor of educational technology at The Open University, UK.
e-mail: mike.sharples@open.ac.uk

Correspondence

Stop government picking professors

A September decree by the Italian government aims to recruit leading university professors through an unprecedented and highly questionable procedure. On behalf of Group 2003 (see www.gruppo2003.org), I urge the Italian government to withdraw these resolutions, which breach the academic freedom of the country's scientists and threaten the future of Italian science.

According to the ruling, the prime minister will appoint the chairs of the recruiting panels in different research areas. Each panel comprises only the chair (nominated by the government) and two other members, both chosen by the chair.

This government-controlled university appointment procedure is intended to replace the peer-review methods that are standard in academia worldwide. To our knowledge, it would be the first such system to operate in a democratic country.

Luigi Nicolais *University of Naples Federico II, Naples, Italy.*
nicolais@unina.it

Trump: renewables for self-sufficiency

US president-elect Donald Trump hopes to achieve energy independence for his country. But even sustaining current energy production will be hard, given that US production of 'tight' oil — extracted from shale rock using fracking — is almost 20% below its March 2015 peak, and shale-gas production is 5% below its February 2016 peak.

In August 2016, the United States imported 8 million barrels of crude oil per day, or 48% of its crude-oil requirements. Tight oil currently accounts for roughly half of US oil, so its production would need to almost triple to replace current imports. This would escalate drilling rates and rapidly exhaust core supply areas, setting the stage for a medium-term

production collapse and radically higher prices. Coupled with a comparably aggressive ramp-up of shale-gas production, this increased activity would compound environmental and human-health risks (see, for example, M. Finkel *Nature* **540**, 39; 2016).

We agree with renewables specialist Daniel Kammen that low-carbon alternatives such as wind and solar are the way to go, particularly because job growth and return on investment should be more robust than those from carbon-based energy (see *Nature* <http://doi.org/bs58>; 2016).

Seth B. C. Shonkoff *PSE Healthy Energy, Oakland, California, USA.*
sshonkoff@psehealthyenergy.org
*On behalf of 7 correspondents (see go.nature.com/2hkumhi for full list).

Trump: time to seize environmental gains

The United States has led the global environmental movement since the 1970s, albeit intermittently. If it withdraws support for multilateral treaties under President Trump, the environment will not be doomed.

China, for example, could step into the lead (see D. Victor *Nature* **539**, 495; 2016). China is committed to the Convention on Biological Diversity, which the United States has still not ratified, and to many other international environmental treaties (F. Wu *J. Chin. Polit. Sci.* **14**, 383–406; 2009). If other countries support China, environmental gains can continue — irrespective of a weakened US contribution.

A Trump government that is less concerned about the environment could create space for strengthened independent initiatives, such as commitments to sustainability, by subnational units of government, cities, companies and community groups (N. Lutsey and D. Sperling *Energy Policy* **36**, 673–685; 2008).

And if Trump's promised trade protectionism occurs, scientists

could help to shape policies that safeguard the environment — such as by restricting imports from regions that do not uphold good environmental practices.

Duan Biggs *Griffith University, Nathan, Queensland, Australia.*
d.biggs@uq.edu.au

*On behalf of 4 correspondents (see go.nature.com/2gfd7ki for full list).

Diversity is future for genetic analysis

The Population Architecture using Genomics and Epidemiology (PAGE) study, to which I contribute, is overcoming some of the technical challenges of multi-ethnic genomic analyses (see A. B. Popejoy and S. M. Fullerton *Nature* **538**, 161–164; 2016). It is yielding findings that are unattainable for homogeneous populations.

PAGE, funded by the US National Institutes of Health, uses genome-wide analyses of 50,000 phenotyped participants of mainly Hispanic and African ancestry. We aggregate results across several studies: for example, two are multi-ethnic, one involves only women and one is designed to address underrepresentation of Hispanic people (see go.nature.com/2hity2j).

We collaborated with other initiatives, including the Consortium on Asthma among African-ancestry Populations in the Americas, to develop an unbiased genotyping array for use across all major continental populations (www.pagestudy.org/mega). Statistical tools such as SUGEN (D. Y. Lin *et al. Am. J. Hum. Genet.* **95**, 675–688; 2014) and GENESIS (M. P. Conomos *et al. Am. J. Hum. Genet.* **98**, 127–148; 2016) can account for the admixed ancestry of individuals (a significant factor in almost every US minority), and for cohorts that include many ancestries.

Leveraging sample diversity in these and other ways has maximized the power of our genetic analyses.

Christopher S. Carlson *Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.*
ccarlson@fhcrc.org

Sustainable fisheries need reserves

Indigenous people such as the Maori and other Polynesians traditionally maintain that the *mauri* (life force) of the ocean must be protected if humans are to prosper. This is echoed by the United Nations Convention on the Law of the Sea, which permits only sustainable fishing activity and aims to protect marine biodiversity. However, far too few governments are honouring that commitment.

Simply making fisheries responsible for conserving nature (R. Hilborn *Nature* **535**, 224–226; 2016) may not work because of the conflict of interest with maximizing catch. In our view, fisheries should instead be under the jurisdiction of agencies whose primary responsibility is to protect biodiversity. Conservation is already using simpler, easy-to-manage and cost-effective methods to protect fisheries, and more than 90% of marine protected areas (MPAs) allow fishing that supports sustainable fisheries (M. J. Costello and B. Ballantine *Trends Ecol. Evol.* **30**, 507–509; 2015).

Integrating fisheries into conservation agencies would prioritize the health of ecosystems, for example by establishing marine reserves (no-take MPAs). Very large examples of such reserves have recently been created by indigenous Pacific islanders in Palau and Kiribati (see also *Nature* **539**, 13–14; 2016). These safe havens can provide stock and spillover for fisheries, as well as baselines for unfished ecosystems and other benefits.

Mark John Costello *University of Auckland, New Zealand.*
m.costello@auckland.ac.nz
*On behalf of 4 correspondents (see go.nature.com/2gollxm for full list).

PHARMACOLOGY

Inside-out receptor inhibition

Structures of two chemokine receptor proteins in complex with small molecules reveal a previously unknown binding pocket that could be a drug target for treating a range of diseases involving this receptor family. [SEE LETTERS P.458 & P.462](#)

THOMAS P. SAKMAR & THOMAS HUBER

A family of cell-membrane proteins known as G-protein-coupled receptors mediates transmembrane signal transduction. One subset of this family is the chemokine receptors, which regulate cell migration and whose activation has been implicated in a range of diseases, including immune disorders and cancer. But finding drugs that inhibit these receptors has been challenging. Two papers in this issue^{1,2} now describe crystal structures of two different chemokine receptors in complex with small-molecule inhibitors. Two of these antagonists bind to pockets near to the receptors' intracellular surfaces, pointing to a previously unidentified pathway that can be targeted for drug discovery.

Most drug molecules that target G-protein-coupled receptors (GPCRs) mimic the binding activity of a native activator, enhancing or inhibiting receptor signalling to achieve a therapeutic effect. The drugs typically occupy a binding pocket called the orthosteric site in the transmembrane region of the receptor that is accessible to the outside of the cell. But the affinity with which activating ligands bind to receptors can be increased by the binding of a G protein. This phenomenon, known as an allosteric effect, is well established in GPCR pharmacology³ and provides an alternative avenue for drug discovery.

Unlike G proteins, which bind to the intracellular side of the receptor, other allosteric molecules tend to bind to sites that are within the membrane region or at the extracellular surface, sometimes even overlapping with the orthosteric pocket. However, a few drug candidates and antibodies seem to bind to the cytoplasmic surface of GPCRs (including chemokine receptors) and affect function^{4–7}. An allosteric drug that binds at the cytoplasmic surface of a GPCR has not been described in detail, until now.

In the first study, Oswald *et al.*¹ (page 462) report the crystal structure of the chemokine receptor CCR9 in complex with a small-molecule drug called vercirnon, which acts as an antagonist to CCR9 activity (Fig. 1a). Inhibition of CCR9 is desirable as a possible way to treat inflammatory bowel disease, but vercirnon did

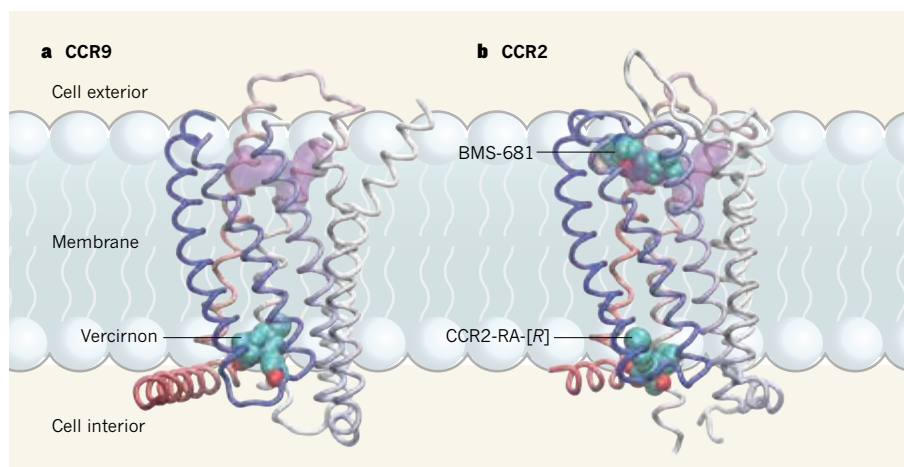


Figure 1 | Visualizing allosteric inhibition. Two papers report structures of chemokine receptor proteins in complex with small-molecule antagonists, which inhibit receptor activity. In both structures, the receptor comprises seven transmembrane helical domains and an eighth helix that lies along the cytoplasmic surface of the membrane (different domains denoted by gradually changing colours, from blue and white to pink). **a**, Oswald *et al.*¹ report that the small molecule vercirnon binds in a pocket on the intracellular side of the CCR9 receptor. The binding pocket for the drug maraviroc in another chemokine receptor, CCR5, is shown in purple for comparison¹². **b**, Zheng *et al.*² report the structure of CCR2 in simultaneous complex with two different antagonists: BMS-681, which binds in a pocket that overlaps with the maraviroc binding site, and CCR2-RA-[R], which binds in an intracellular pocket almost identical to that in CCR9.

not pass a phase III trial in people with Crohn's disease. Crystallization of the CCR9-vercirnon complex required the use of a CCR9 variant that had eight amino-acid substitutions and truncated amino- and carboxy-terminal tails, but the authors confirmed that none of these changes affected vercirnon binding.

The structure reveals the seven transmembrane helical segments of CCR9 connected by three cytoplasmic loops, with an eighth helix that seems to rest on the intracellular surface of the membrane. Vercirnon is an asymmetric, inverted-V-shaped structure that binds in a pocket comprising five of the seven helices, and it peeks out directly into the cytoplasm. The vercirnon binding site is about 33 ångströms from the presumed orthosteric pocket, which lies towards the extracellular surface of the seven-helical bundle.

In the second study, Zheng *et al.*² (page 458) report the structure of CCR2 — which has been implicated in various chronic inflammatory and autoimmune disorders and in anti-tumour immunity — in simultaneous complex

with two small-molecule antagonists dubbed CCR2-RA-[R] and BMS-681 (Fig. 1b). The authors enabled crystallization by truncating the carboxy-terminal tail of CCR2, and by fusing a stabilizing protein called T4 lysozyme into an altered third cytoplasmic loop, which is a common strategy to facilitate GPCR crystallization.

BMS-681 binds in a pocket that overlaps with the presumed orthosteric binding site near the extracellular surface of CCR2, whereas CCR2-RA-[R] binds in a site remarkably like the allosteric pocket in CCR9. The simultaneous binding of the two antagonists traps CCR2 in a conformation that seems to be completely inactive, on the basis of the tight helical arrangement of the protein and the absence of regional conformations characteristic of receptor activation. The helix bundle looks similar to that of rhodopsin, a light-activated GPCR, when the protein is in the dark — the gold standard for an inactive GPCR structure.

Although vercirnon and CCR2-RA-[R] are different chemical entities, they occupy binding pockets that lie in the same location and

have a three-dimensional lining made up of amino-acid side chains from helices 1, 2, 3, 6 and 7. These similarities indicate that the allosteric pocket might be present in other chemokine receptors. If this supposition holds true, medicinal and computational chemists will have a field day using structure-aided design strategies to develop drugs that target the pocket. Vercirnon, for example, was not optimized for the CCR9 pocket, and it is likely that some minor alterations would improve its drug properties.

The structures also provide insights into the mechanism of intracellular allosteric antagonism. Both show that the bound antagonists prevent outward movement and rotation of helices (especially helix 6), which is the hallmark of the active-state structure. Particularly in the CCR9-vercirnon structure, in which the cytoplasmic loops are not modified for crystallization and are reasonably well resolved, it is clear that vercirnon occupies a space that would normally be filled by the carboxy-terminal tail of a bound G protein during receptor activation⁸. Binding by the protein β -arrestin, which inhibits signalling and causes receptors to be internalized into the cell, would also certainly clash with bound vercirnon⁹.

Although structures have been produced for the human chemokine receptor CXCR4 in complex with a chemokine from a virus¹⁰, and for a viral chemokine receptor in complex with the human chemokine CX3CL1 (ref. 7), there is not yet a structure of a human chemokine receptor in complex with a human chemokine. Chemokines themselves are relatively complex protein ligands, generally comprising about 70 amino acids. What is clear is that multiple regions of the extracellular surface of chemokine receptors have roles in docking the chemokine before it can engage the orthosteric pocket in the helix bundle¹¹. Identification of a cytoplasmic allosteric binding pocket in chemokine receptors is especially valuable, because it provides an alternative strategy for structure-based drug discovery before the precise binding mode of chemokines has been fully elucidated.

Progress in developing useful small-molecule drugs for chemokine receptors has been slow. There have been just two successes — maraviroc, which targets CCR5 to prevent HIV-1 from entering cells, and plerixafor, which targets CXCR4 to mobilize bone-marrow stem cells for transplantation in people with cancer. Another dozen or so chemokine receptors are drug targets for diseases ranging from autoimmune disorders to cancer metastasis. The intracellular binding pocket identified in the current studies might provide a new strategy for inhibiting these receptors, by turning drug-discovery efforts inside out. ■

Thomas P. Sakmar and Thomas Huber
are in the Laboratory of Chemical Biology

and Signal Transduction, The Rockefeller University, New York, New York 10065, USA.
e-mails: sakmar@rockefeller.edu;
hubert@rockefeller.edu

1. Oswald, C. *et al.* *Nature* **540**, 462–465 (2016).
2. Zheng, Y. *et al.* *Nature* **540**, 458–461 (2016).
3. De Lean, A., Stadel, J. M. & Lefkowitz, R. J. *J. Biol. Chem.* **255**, 7108–7117 (1980).
4. Zweemer, A. J. M. *et al.* *Mol. Pharmacol.* **84**, 551–561 (2013).
5. Tchernychev, B. *et al.* *Proc. Natl Acad. Sci. USA* **107**,

- 22255–22259 (2010).
6. Staus, D. P. *et al.* *Nature* **535**, 448–452 (2016).
7. Burg, J. S. *et al.* *Science* **347**, 1113–1117 (2015).
8. Rasmussen, S. G. F. *et al.* *Nature* **477**, 549–555 (2011).
9. Kang, T. *et al.* *Nature* **523**, 561–567 (2015).
10. Qin, L. *et al.* *Science* **347**, 1117–1122 (2015).
11. Veldkamp, C. T. *et al.* *Sci. Signal.* **1**, ra4 (2008).
12. Tan, Q. *et al.* *Science* **341**, 1387–1390 (2013).

This article was published online on 7 December 2016.

CHEMICAL BIOLOGY

A radical change in enzyme catalysis

A method has been devised that allows a ketoreductase enzyme to catalyse reactions other than its natural ones. The key is to excite the enzyme's cofactor using light — an approach that might work for other enzymes. SEE LETTER P.414

UWE T. BORNSCHEUER

Enzymes have several advantages over conventional catalysts for organic synthesis. For example, their ability to perform reactions at room temperature in water makes them suitable for environmentally friendly chemical processes. But many synthetically useful reactions cannot be catalysed by naturally occurring enzymes. The quest to expand nature's enzymatic repertoire of transformations is therefore a crucial area of research. On page 414, Emmanuel *et al.*¹ report a strategy that allows ketoreductase enzymes to perform completely different reactions from the ones that they evolved to catalyse.

Living organisms are without doubt the best chemists on Earth — a plethora of reaction types is catalysed by the thousands of different enzymes present in every cell. The reactions take place with excellent selectivity (forming solely the desired product), astonishing efficiency (performing hundreds of catalytic reactions per second at a single catalytic site) and at ambient temperatures and pH values. By contrast, chemists have developed methods that allow a range of reactions with no enzymatic counterpart to be easily performed. In many cases, developing an enzyme that can perform such reactions is desirable.

Researchers have therefore devised a range of concepts for creating or modifying proteins to catalyse reactions unknown in nature^{2–4}. One approach is to incorporate chemical transition-metal catalysts into protein scaffolds. As early as 1978, the protein avidin was modified to incorporate a rhodium catalyst⁵, producing an enzyme that catalyses asymmetric hydrogenations — transformations in which hydrogen reacts

with organic molecules to produce products predominantly as one mirror-image isomer (enantiomer). This year, a system in which a ruthenium catalyst was incorporated into the streptavidin protein enabled olefin metathesis (a carbon–carbon bond-formation reaction; Fig. 1a) *in vivo* in the bacterium *Escherichia coli*⁶.

A second general approach is to use protein engineering to redesign enzymes to catalyse reactions other than the native one. This strategy is exemplified by the engineering of P450 monooxygenase enzymes⁷ to catalyse carbon–carbon bond-formation reactions (cyclopropanations; Fig. 1b), rather than the analogous carbon–oxygen bond-formation reactions (epoxidations) that occur naturally. A third approach is computational *de novo* protein design, which has been used to make an enzyme that catalyses the Kemp elimination reaction, in which a hydrogen ion (H⁺) is removed from a carbon atom in an organic molecule (Fig. 1c)⁸. Subsequent extensive protein engineering through directed evolution of the enzyme resulted in catalytically efficient mutants⁹.

Emmanuel *et al.* now report a striking new concept for generating enzymes that catalyse unnatural reactions: the authors use light to excite a cofactor (NAD(P)H) bound in the active site of a ketoreductase (KRED) enzyme. The resulting photoexcited cofactor generates a radical intermediate that serves as a hydrogen source. Furthermore, this hydrogen source is chiral — it has a 'handedness' that can potentially be passed on to other molecules during reactions. The authors find that, when KRED contains a photoexcited cofactor, it catalyses a reaction in which a halogen atom is removed from molecules known as halolactones,

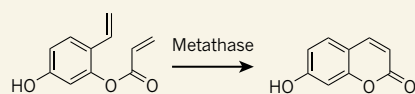
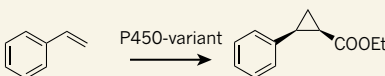
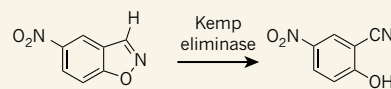
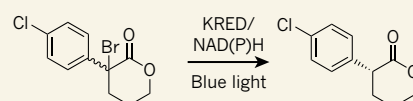
a Olefin metathesis**b Cyclopropanation****c Kemp elimination****d Radical-induced dehalogenation**

Figure 1 | Non-natural reactions catalysed by enzymes. Several strategies have been devised to develop enzymes that catalyse reactions unknown in nature. **a**, Synthetic catalysts can be incorporated into naturally occurring proteins. This approach was used to prepare metathase enzymes that catalyse olefin metathesis⁶. **b**, Enzymes can be engineered to catalyse non-natural reactions — for example, P450 enzymes have been engineered to promote the cyclopropanation reaction⁷. **c**, Kemp elimination enzymes have been designed *de novo* computationally⁸, to perform the Kemp elimination reaction. **d**, Emmanuel *et al.*¹ now report that the cofactor NAD(P)H can be excited by blue light in ketoreductase (KRED) enzymes. This enables the enzyme to catalyse an unnatural reaction known as radical-induced dehalogenation, which yields the product as predominantly one mirror-image isomer (enantiomer). Et, ethyl; Cl, chlorine; Br, bromine.

forming products predominantly as one enantiomer (Fig. 1d). Moreover, the enantiomer that is formed depends on the preference of the KRED that is used. The authors show that this unnatural reaction can be used to generate either of the enantiomers of products formed from a broad range of halolactones, demonstrating the synthetic usefulness of this approach.

The KRED fulfils two functions in this reaction. First, it ensures productive, coordinated binding of the photoexcited NAD(P)H with the halolactone in its active site. But it also recycles the spent cofactor by reacting it with isopropanol (a component of the reaction mixture), regenerating NAD(P)H. This efficient recycling enables a KRED molecule to mediate multiple catalytic cycles, as would be needed for the enzyme to be used to make gram or kilogram quantities of product for industrial applications.

Not all the KREDs investigated by the authors could catalyse the reaction; Emmanuel and colleagues found that certain point mutations in the enzyme are needed to promote the productive binding of NAD(P)H within the enzyme's scaffold. However, the catalytically active KREDs bind the halolactones perfectly, even though they do not resemble the enzymes' natural substrates. Furthermore, the authors proved that the unnatural reaction occurs only when NAD(P)H is tightly bound to KRED and is irradiated with blue light.

The authors proposed a mechanism for the reaction in which light irradiation causes an electron to be transferred between the NAD(P)H and the substrate, triggering cleavage of the substrate's carbon–halogen bond,

and thus generating a radical intermediate that accepts a hydrogen atom to form the final product enantioselectively (see Fig. 3d of the paper¹). They nicely confirm this mechanism by generating a deuterium donor from NAD(P)H *in situ* in KRED, and observing where the deuterium is incorporated into the reaction products.

Emmanuel *et al.* have demonstrated a completely new strategy for accessing

unnatural enzymatic reactions by exploring the interface between photochemistry and protein science. Other synthetic transformations can be envisaged with this approach, by using light-induced changes in NAD(P)H analogues or other cofactors. For instance, the well-studied flavin cofactors (flavin adenine dinucleotide, flavin mononucleotide and their artificial analogues) could be prime candidates for investigation, because various flavin-dependent enzymes are important biological catalysts used by synthetic chemists¹⁰. In combination with modern tools for protein engineering¹¹, the authors' concept is likely to have a strong impact on the use of various enzyme classes in biocatalysis. ■

Uwe T. Bornscheuer is in the Department of Biotechnology and Enzyme Catalysis, Institute of Biochemistry, Greifswald University, 17489 Greifswald, Germany.

e-mail: uwe.bornscheuer@uni-greifswald.de

- Emmanuel, M. A., Greenberg, N. R., Oblinsky, D. G. & Hyster, T. K. *Nature* **540**, 414–417 (2016).
- Hyster, T. K. & Ward, T. R. *Angew. Chem. Int. Edn* **55**, 7344–7357 (2016).
- Renata, H., Wang, Z. J. & Arnold, F. H. *Angew. Chem. Int. Edn* **54**, 3351–3367 (2015).
- Bornscheuer, U. T. *et al. Nature* **485**, 185–194 (2012).
- Wilson, M. E. & Whitesides, G. M. *J. Am. Chem. Soc.* **100**, 306–307 (1978).
- Jeschek, M. *et al. Nature* **537**, 661–665 (2016).
- Coelho, P. S., Brustad, E. M., Kannan, A. & Arnold, F. H. *Science* **339**, 307–310 (2013).
- Röthlisberger, D. *et al. Nature* **453**, 190–195 (2008).
- Blomberg, R. *et al. Nature* **503**, 418–421 (2013).
- Toogood, H. S., Gardiner, J. M. & Scrutton, N. S. *ChemCatChem* **2**, 892–914 (2010).
- Kazlauskas, R. J. & Bornscheuer, U. T. *Nature Chem. Biol.* **5**, 526–529 (2009).

CANCER

A gene-expression profile for leukaemia

Can simple genetic risk profiles be identified for complex diseases? The development of a gene-expression profile for acute myeloid leukaemia suggests that they can, and that they may improve prognosis prediction. SEE LETTER P.433

GERRIT J. SCHUURHUIS

On page 433, Ng *et al.*¹ report a tool that improves the prediction of prognoses for people who have a form of acute leukaemia. The researchers began by identifying populations of cells that exhibit key properties — collectively known as stemness — that enable the cells to initiate and sustain leukaemia. This allowed the authors to ascertain gene-expression profiles for stemness, and to use them as the basis of a scoring system for risk. The work demonstrates how

gene-expression profiles can be used to enable reliable prognoses for complex diseases.

Acute myeloid leukaemia (AML) is characterized by the presence of a huge range of chromosomal and molecular aberrations. This means that there are many subgroups of people with AML who have widely different prognoses². These groups are known as risk groups, and are used to determine which consolidation treatment should be given following initial chemotherapy (induction therapy). For example, transplantation of stem cells from donors is an option for patients judged to be

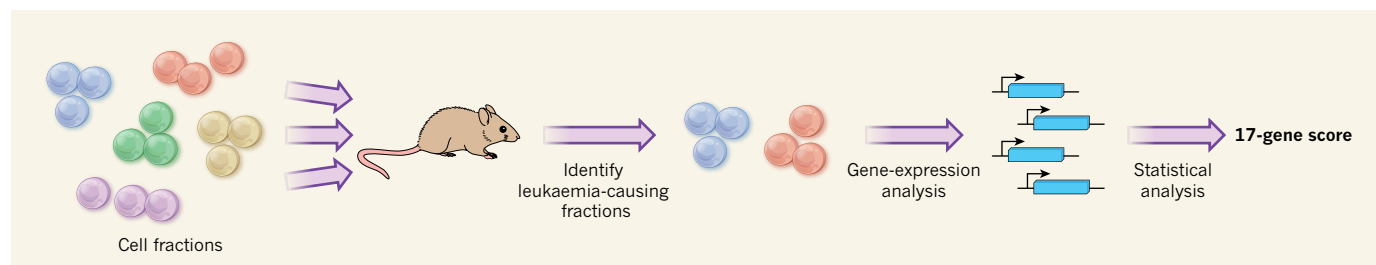


Figure 1 | A 17-gene score for assessing the risk of acute leukaemia.

Ng *et al.*¹ took cell samples from people with acute myeloid leukaemia (AML) and divided them into fractions based on the expression of CD34 and CD38 proteins on the cells' surfaces. The researchers transplanted the fractions into mice, and identified which fractions caused leukaemia and which did not. The authors then compared gene-expression patterns in the disease-causing cell

fractions with those in the non-disease-causing fractions, and thus identified candidate genes that correlated with tumour formation. This information was used to direct a statistical analysis of gene-expression data that had previously been gathered in a clinical study⁷ of people with AML. The analysis identified a score that could be calculated for patients based on the expression of 17 genes. The score provides a reliable system for assessing patients' prognosis.

at the most risk from the disease. But such transplantation often has fatal side effects³, so is not the best choice for some patients. Improvements to risk assessments are necessary, not only to make decisions about consolidation strategies, but also to choose between different types of induction therapy (which are expected to become available in the future).

Gene-expression profiles could be instrumental in realizing these improvements⁴. Ng and colleagues' approach, which relies on identifying profiles for stemness⁵, is a good example of how such profiles could be used. In normal tissue, stemness allows stem cells to self-renew — to sustain the long-term process of normal cell differentiation. In the haematopoietic system, normal haematopoietic stem cells (HSCs) are the origin of blood cells in the circulation and the bone marrow. HSCs express CD34 proteins on their surfaces, but not CD38 proteins, and are thus said to have the CD34⁺CD38⁻ immunophenotype. Leukaemic stem cells (LSCs) have stemness properties similar to those of HSCs, but they can express different patterns of cell-surface proteins: they can be CD34⁺CD38⁻ cells (which are probably derived from HSCs), but they can also have CD34⁺CD38⁺, CD34⁻CD38⁺ or CD34⁻CD38⁻ immunophenotypes. It has previously been shown in animal models that the leukaemia-initiating ability of these different CD34/CD38 subpopulations can differ⁶.

In a huge effort, Ng *et al.* isolated 227 CD34/CD38-defined cell fractions from 78 people with AML, and injected the fractions into mice (Fig. 1). They confirmed that the leukaemia-initiating ability of the cell fractions differed: leukaemia could form from all the cell fractions obtained from a patient, from some of the fractions or from none. The authors then compared gene expression in the original cell fractions that caused leukaemia with gene expression in cell fractions that did not, irrespective of the cells' CD34/CD38 immunophenotype. This allowed them to identify gene-expression patterns that were directly related to the ability of cells to form leukaemias *in vivo* in mice.

Ng and colleagues first identified 104 genes

for which expression levels differed by at least twofold in leukaemia-initiating cell fractions compared with fractions that didn't initiate leukaemia. The authors then examined a large set of gene-expression data obtained from a clinical study⁷ of 495 people with AML, and found that 89 of the 104 genes were present in the set. The cells in that study were not divided into fractions, but displayed gene-expression patterns that were similar to those observed by Ng *et al.* in leukaemia-initiating cell fractions.

Next, the authors used a statistical method to relate gene expression to clinical outcome for these 89 genes, and for a subset of 43 genes that are highly expressed in leukaemia-initiating cell fractions. This allowed them to identify an optimal panel of 17 genes, the expression of which was highly indicative of a poor clinical outcome in a patient subgroup. The authors confirmed this finding in other AML cohorts and found that a scoring system based on their gene panel (called LSC17) offered superior prognoses when compared with other gene-expression profiling systems for AML^{5,8}. In fact, Ng and colleagues found that previously reported genetic signatures of AML were not independent prognostic factors when tested in the other cohorts.

Ng *et al.* also found that gene-expression patterns associated with stemness in AML are independent of the chromosomal and molecular aberrations used to assess patient risk, showing that stemness is a factor that crosses the borders of previously identified risk groups. Finally, the authors developed an assay that allows gene-expression data to be rapidly generated, which could form the basis of a fast (24–48 hours) prognostic test for patients.

As the authors indicate, analysis of large data sets from clinical studies in which both extensive information about the mutational status of leukaemia cells⁹ and LSC17 scores are available will be needed to assess whether the prognostic value of the LSC17 score is independent of the prognostic value of mutations present at diagnosis. The clinical benefits of the LSC17 score must be assessed, because prognostic

value does not always lead to a meaningful clinical advantage. Moreover, small populations of leukaemia cells that have a similar genetic make-up (clones) can be present at diagnosis, survive therapy and proliferate to cause a relapse (in some cases after having acquired additional mutations^{10,11}). Only time will tell whether Ng and co-workers' gene-expression profiles account for cell fractions defined by such clones, and thereby predict associated relapses.

The prognosis of a person with leukaemia at the point of diagnosis is only part of the prognostic story. Once treatment has started, factors such as therapy compliance, alterations to drug doses that are made to mitigate side effects, and differences between patients in the concentration of drugs in blood plasma might partly override the effects of prognostic diagnosis parameters such as gene-expression patterns. The consideration of post-treatment parameters such as measurable (minimal) residual disease¹² (a measure of the persistence of small numbers of leukaemia cells in patients in remission) has drastically changed the landscape of risk assessment in AML. Assessing combinations of cellular properties at diagnosis, non-cellular patient-specific factors during therapy, frequencies and properties of cells that remain after treatment and changes in immunological parameters, might offer a more-refined prognosis than is currently possible. This would enable more-personalized induction and consolidation treatments to be used. Ng and colleagues' study is potentially a big step towards such assessments, especially at the diagnosis stage. ■

Gerrit J. Schuurhuis is in the Department of Hematology, VU University Medical Center, De Boelelaan 1117, 1081HV Amsterdam, the Netherlands.
e-mail: gj.schuurhuis@vumc.nl

1. Ng, S. W. K. *et al.* *Nature* **540**, 433–437 (2016).
2. Grimwade, D. *et al.* *Blood* **116**, 354–365 (2010).
3. Cornelissen, J. J. *et al.* *Nature Rev. Clin. Oncol.* **9**, 579–590 (2012).
4. Shivarov, V. & Bullinger, L. *Exp. Hematol.* **42**, 651–660 (2014).
5. Levine, J. H. *et al.* *Cell* **162**, 184–197 (2015).

6. Sarry, J.-E. *et al.* *J. Clin. Invest.* **121**, 384–395 (2011).
 7. Verhaak, R. G. W. *et al.* *Haematologica* **94**, 131–134 (2009).
 8. Gentles, A. J. *et al.* *J. Am. Med. Assoc.* **304**,

- 2706–2715 (2010).
 9. Papaemmanuil, E. *et al.* *N. Engl. J. Med.* **374**, 2209–2221 (2016).
 10. Ding, L. *et al.* *Nature* **481**, 506–510 (2012).

11. Jan, M. & Majeti, R. *Oncogene* **32**, 135–140 (2013).
 12. Hokland P. *et al.* *Semin. Hematol.* **52**, 184–192 (2015).

This article was published online on 7 December 2016.

HYDROLOGY

The dynamics of Earth's surface water

High-resolution satellite mapping of Earth's surface water during the past 32 years reveals changes in the planet's water systems, including the influence of natural cycles and human activities. SEE LETTER P.418

DAI YAMAZAKI & MARK A. TRIGG

Everyone appreciates that the water cycle can vary, and can cause floods and droughts at its extremes. On page 418, Pekel *et al.*¹ map the full range of this variability, as evidenced by our rivers, lakes and wetlands, using more than 3 million satellite images collected over the past 32 years. This globally consistent analysis documents both natural water variability and humanity's major influence on Earth's water systems, and will provide a valuable baseline for observations of the effects of future climate change.

Detailed maps describing the location and extent of rivers, lakes and wetlands are needed for many studies of Earth science, but the full global distribution and variability of these systems has not been clearly understood. Scientists have developed methods to map water bodies using satellite observations — for example, by detecting the characteristic reflectance of sunlight from water. But this is a particularly challenging task because the colour of water varies greatly depending on depth, the presence of suspended sediments and dissolved chemicals, and the angle at which sunlight hits the surface. In addition, some land surfaces (such as snow, ice and lava) have similar reflectance characteristics to those of water, which means that water-detection algorithms need to be developed and calibrated carefully.

The first global surface-water map² made using satellite observations was developed in 2009, but computational power restricted the spatial resolution to 250 metres, which is too low to enable detailed mapping of smaller lakes and rivers. This was a problem, because statistical estimates³ suggest that millions of lakes less than 1 square kilometre in size could account for about 40% of the global area of inland water. The situation has since improved: a global analysis of water bodies at 30-metre resolution was undertaken recently^{4,5} using images from the Landsat programme (the world's longest-running initiative for acquiring satellite images of Earth).

However, the location and extent of water bodies can change with time, in part because of natural processes such as flooding, sedimentation and channel migration, but also because of human processes such as dam construction and water abstraction. This creates a need for a global-scale, high-resolution analysis of information taken at different times — a complete map of surface-water dynamics. Such dynamics have recently been captured in maps that enable scientists to distinguish permanent rivers and lakes from seasonal water bodies such as flood plains⁶ and to explore the long-term trends of surface-water changes⁷, but these studies used only a subset of all the Landsat images available.

Pekel and colleagues' ambitious work uses the entire Landsat archive⁸ to map global surface waters — more than 3 million images collected between 1984 and 2015. To handle this petabyte-scale data set, the authors used Google Earth Engine (go.nature.com/2fdt80k), a freely available cloud-computing platform for analysing big data sets of satellite observations. The Landsat data set was produced

using three satellites, and multiple operational issues affected the collection and quality of the data. This presented unique challenges, in addition to those associated with water's variable reflective properties. To overcome these challenges, Pekel *et al.* used a combination of expert systems (computer systems that use artificial intelligence) and visual analytics to identify the existence or absence of surface water for every pixel of Earth imagery, each representing a square of side 30 metres; this was done at monthly intervals over the 32-year period.

An understanding of the frequency with which water occurs at different locations is certainly a useful result of such an analysis. However, more-meaningful information and visualization of global-scale changes are required to cope with gaps in the data set that result from cloud cover and operational deficiencies, and to allow specific interpretation of different surface-water dynamics such as seasonal cycle and long-term trend. Pekel and colleagues therefore provide thematic maps depicting persistence (whether water is always present, or just sometimes), gains versus losses, the consistency of seasonal cycles, permanent versus seasonal water, and transitions between seasonal and permanent water during the period analysed (Fig. 1). The output of the analysis and the thematic maps are available through a user-friendly interface (go.nature.com/2gj81ap), allowing anyone to explore any location and understand what surface-water changes have occurred, without the need for complex analysis or massive computing power.

The authors' high-quality analyses and visualizations of the data reveal that there were 2.78 million km² of permanent

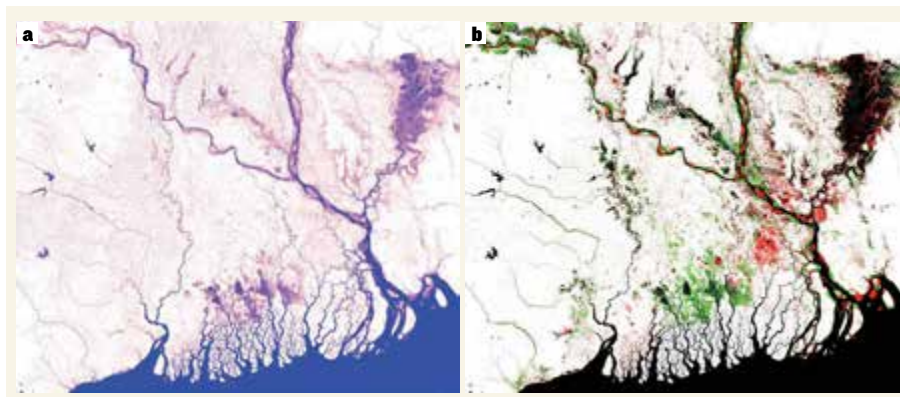


Figure 1 | Maps showing variability of surface water in the Ganges delta. Pekel *et al.*¹ have used historical satellite images to produce global maps that depict changes of surface water over the past 32 years. The maps are presented in different ways to enable different information to be visualized. **a**, This map shows the average water-occurrence frequency over 32 years; blue represents water that is always there, pink is water that is sometimes there. **b**, Here, red regions indicate where water occurrence has decreased during the period studied, whereas green indicates increased occurrence. These maps, along with others that depict seasonal variations, help to distinguish different causes of water dynamics, such as seasonal inundation, channel migration and reservoir construction.

J.-F. PEKEL ET AL./REF. 1

surface water and 0.81 million km² of seasonal surface water on Earth in 2015. During the full period of the analysis, 162,000 km² of permanent water were lost or became seasonal, whereas 184,000 km² of new permanent waters were created at different locations. More than 70% of the losses were concentrated in just five countries (Kazakhstan, Uzbekistan, Iran, Iraq and Afghanistan) clustered in the Middle East and Asia, raising serious questions about water security and transboundary water management in that region. Most of the permanent-water gain correlates with reservoir construction worldwide, but the impact of climate change was also detected in lake expansion caused by melting glaciers in the Tibetan Plateau. Changes that occur across decades, such as those due to the recent drought in Australia, also stand out clearly.

Any analysis that quantifies surface water from historical data sets will have limitations. In this case, data gaps affect the accuracy of the seasonality information; resolution issues prevent analysis of small water bodies; vegetation obscures important wetlands; and the 16-day repeat cycle of Landsat observation means that events that occur on shorter timescales, such as floods, may be missed. These problems will be addressed in the future by using better optical and radar sensors and more satellites, and by integrating satellite-observed data into models of surface-water dynamics.

Despite the limitations, Pekel *et al.* have provided our best understanding yet of the changes in our planet's surface water. Their findings will be crucial to many Earth-science studies — such as climate-modelling efforts, or investigations of ecology at the interfaces between land and rivers — and for global water-management initiatives. ■

Dai Yamazaki is in the Department of Integrated Climate Projection Research, Japan Agency for Marine–Earth Science and Technology, Yokohama, Kanagawa 237-0061, Japan. **Mark A. Trigg** is in the School of Civil Engineering, University of Leeds, Leeds LS2 9JT, UK.
e-mails: d-yamazaki@jamstec.go.jp; m.trigg@leeds.ac.uk

1. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. *Nature* **540**, 418–422 (2016).
2. Carroll, M. L., Townshend, J. R., DiMiceli, C. M., Noolipady, P. & Sohlberg, R. A. *Int. J. Dig. Earth* **2**, 291–308 (2009).
3. Downing, J. A. *et al.* *Limnol. Oceanogr.* **51**, 2388–2397 (2006).
4. Verpoorter, C., Kutser, T., Seekell, D. A. & Tranvik, L. J. *Geophys. Res. Lett.* **41**, 6396–6402 (2014).
5. Feng, M., Sexton, J. O., Channan, S. & Townshend, J. R. *Int. J. Dig. Earth* **9**, 113–133 (2016).
6. Yamazaki, D., Trigg, M. A., Ikeshima, D. *Remote Sens. Environ.* **171**, 337–351 (2015).
7. Donchyts, G. *et al.* *Nature Clim. Change* **6**, 810–813 (2016).
8. Wulder, M. A. *et al.* *Remote Sens. Environ.* **185**, 271–283 (2016).

This article was published online on 7 December 2016.

STEM CELLS

Cause and consequence in aged-muscle decline

Activation of aged muscle stem cells induces changes in DNA packaging that lead to expression of the gene *Hoxa9*. This reactivates embryonic signalling pathways, restricting the cells' ability to repair injured muscle. SEE LETTER P.428

SUSAN ELIAZER & ANDREW S. BRACK

As muscle stem cells age, their ability to regenerate skeletal muscle following injury declines. One factor that might be responsible is alteration of the highly compact nuclear complex called chromatin, in which DNA is packaged around histone proteins. A changing environment can induce molecular modifications, known as epigenetic changes, to histone proteins, altering chromatin state and so modifying the cell's transcriptional landscape¹ — a more open conformation permits gene transcription, whereas tighter packaging is repressive.

Ageing muscle stem cells exhibit epigenetic alterations², but whether these changes cause the regenerative decline of skeletal muscle with age has been unknown. Schwörer *et al.*³ report on page 428 that the gene *Hoxa9* acts as a molecular node for this dysfunction, being aberrantly expressed as a result of epigenetic modifications in aged muscle stem cells, and in turn promoting abnormal expression of downstream signalling pathways that drive the cells' functional decline.

The authors isolated quiescent stem cells from normal muscle and activated stem cells from injured muscle in young-adult and aged mice, and analysed histone modifications using a strategy based on mass spectrometry. Compared with quiescent stem cells in the young adults, the aged quiescent cells showed increased levels of molecular modifications associated with gene repression (the addition of two methyl groups to the amino-acid residue lysine 9 (K9) on histone H3, and trimethylation of K27 on H3) and lower levels of marks associated with activation (the addition of acetyl groups on H3 and H4, and dimethylation of K36 on H3). The researchers also noticed that the transition from quiescence to activation was accompanied by a decrease in active marks in young-adult mice. By contrast, the transition in aged stem cells was associated with an increase in active marks and a decrease in repressive marks, resulting in a more permissive chromatin state and aberrant gene expression (Fig. 1).

Schwörer and colleagues also observed increased *Hoxa9* transcript and protein levels in activated aged stem cells compared

with those of young adults. This increase was induced by recruitment of an enzyme called Mll1 to the *Hoxa9* region. The enzyme deposits an activation-associated modification on H3 (trimethylation of K4; a mark dubbed H3K4me3). Wdr5, a scaffold protein for Mll1, was also recruited. Together, the two recruited factors caused *Hoxa9* to be transcribed and translated at a much higher level.

The authors then used several strategies to investigate the potential role of *Hoxa9* in the regenerative dysfunction of aged muscle stem cells — mutating *Hoxa9* in mice, or inhibiting *Hoxa9* transcription in muscle- or stem-cell cultures that were transplanted into an injured muscle in young-adult mice. In all cases, *Hoxa9* inhibition restored proliferation and regeneration in aged stem cells. Furthermore, overexpression of *Hoxa9* in young-adult stem cells suppressed their proliferative capability, thus mimicking the aged situation. Together, these data imply that *Hoxa9* expression causes regenerative decline in ageing muscle.

Hox genes are a vital part of embryonic development, determining the anatomical identity of each segment along the head–tail axis of the growing fetus^{4,5}. By investigating signalling pathways known to be regulated by Hox genes during development, Schwörer *et al.* found that overexpression of *Hoxa9* led to the aberrant expression of various developmental pathways, including Wnt, BMP–TGF- β and JAK–STAT — all of which have been shown to alter muscle stem-cell function during ageing^{6–9}.

Although it seems that increased *Hoxa9* expression is the nodal point for aberrant signalling in aged stem cells, inhibiting any one of these signalling pathways can restore stem-cell capacity^{6–9}. This suggests either that there is crosstalk between the signalling pathways, or that the pathways are induced as a consequence of the stress caused by injury and lowering the levels of one pathway mitigates the stress response. It would be interesting to determine how patterns of epigenetic modifications are changed when these signalling pathways are inhibited. Could reducing the level of a single aberrant pathway following injury restore histone-modification patterns to the young-adult state, implying that a feedback loop controls the regenerative capacity of stem cells?

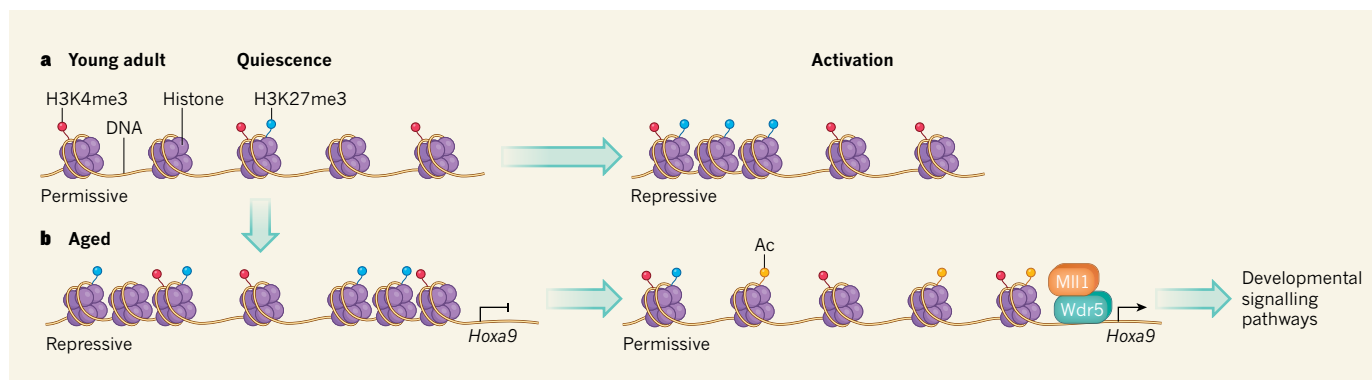


Figure 1 | A changing epigenetic landscape during ageing. Molecular modifications to histone proteins, around which DNA is wrapped as chromatin, determine whether chromatin is loosely packaged — a state permissive for gene transcription — or closed in a repressive state. **a**, In quiescent young-adult-mouse muscle stem cells, chromatin has high levels of activating modifications (for example, H3K4me3) and low levels of repressive marks (H3K27me3), leading to a permissive state. Following injury, when stem cells are activated, chromatin accumulates H3K27me3 and becomes repressive².

b, Schwörer *et al.*³ analysed histone modifications and *Hoxa9* gene expression in aged-mouse muscle stem cells. Aged quiescent stem cells acquire more repressive marks than their young-adult counterparts, presumably repressing *Hoxa9* transcription. Injury results in a global increase of another active mark (acetylation, Ac) and a decline in repressive marks, opening chromatin. There is also an increased recruitment of a Mll1–Wdr5 protein complex to the *Hoxa9* region, which promotes expression of *Hoxa9* and downstream developmental signalling pathways, contributing to the loss of muscle stem-cell function.

Schwörer *et al.* focused on changes in H3K4me3 to explain the transcriptional and subsequent translational difference in *Hoxa9* between activated adult and aged stem cells, but other possibilities cannot be ruled out. RNA sequencing by the authors indicated that *Hoxa9* transcripts are expressed in both adult and aged activated stem cells. However, H3K4me3 is not present at the *Hoxa9* region in activated adult stem cells. Perhaps, then, a different epigenetic mechanism enables *Hoxa9* transcription in adult stem cells, but protein activity is repressed. Alternatively, maybe a subset of activated adult stem cells is in a permissive state, allowing low-level *Hoxa9* gene expression.

It remains unclear why chromatin unwinds in aged stem cells following injury. It is possible that changes in the activity of anti- and pro-ageing factors over a lifetime of wear and tear drive a maladaptive epigenetic response to injury. Alternatively, the chromatin might open as an injury-response mechanism to facilitate DNA repair.

This impressive study demonstrates how abnormal stress-induced epigenetic activation can alter stem-cell function during ageing. The work could have broad medical implications if *Hoxa9* is confirmed to be an intermediary between the epigenetic response to injury and developmental signalling pathways during regeneration in elderly humans. Understanding the upstream events that cause epigenetic de-repression of *Hoxa9* might then be beneficial for strategies to prevent — or even reverse — age-associated declines in muscle regeneration. ■

Susan Eliazer and Andrew S. Brack are in the Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, and the Department of Orthopaedic Surgery, University of California, San Francisco,

San Francisco, California 94143, USA.
e-mail: andrew.brack@ucsf.edu

1. Zhu, J. *et al.* *Cell* **152**, 642–654 (2013).
2. Liu, L. *et al.* *Cell Rep.* **4**, 189–204 (2013).
3. Schwörer, S. *et al.* *Nature* **540**, 428–432 (2016).
4. Pearson, J. C., Lemons, D. & McGinnis, W. *Nature Rev. Genet.* **6**, 893–904 (2005).
5. McGinnis, W. & Krumlauf, R. *Cell* **68**, 283–302 (1992).

6. Brack, A. S. *et al.* *Science* **317**, 807–810 (2007).
7. Carlson, M. E. *et al.* *Aging Cell* **8**, 676–689 (2009).
8. Price, F. D. *et al.* *Nature Med.* **20**, 1174–1181 (2014).
9. Tierney, M. T. *et al.* *Nature Med.* **20**, 1182–1186 (2014).

This article was published online on 30 November 2016.

BIOMEDICINE

An eye on retinal recovery

Retinal-cell transplants restore vision in mouse models of retinal degeneration. It emerges that the transplant leads to an exchange of material between donor and host cells — not to donor-cell integration into the retina, as had been presumed.

MICHAEL A. DYER

The degeneration of the light-sensing photoreceptor neurons in the retina at the back of the eye is a cause of blindness in millions of people worldwide. One possible therapeutic approach to treating advanced photoreceptor degeneration would be to transplant healthy photoreceptor precursors into the damaged eye, in the hope that they would integrate into the retina and restore vision. Such a strategy was shown to be promising in mouse models^{1,2}, raising the possibility of clinical trials in the near future. But three studies^{3–5} in *Nature Communications* report that the transplanted cells rarely integrate into the retina as previously believed — instead, they transfer some of their contents to recipient photoreceptors. This finding is a setback

for efforts to replace lost photoreceptors, but might point to fresh approaches to rejuvenating aged or diseased photoreceptors and decelerating retinal degeneration.

In 2006, a group of researchers transplanted photoreceptor precursors that were indelibly marked with green fluorescent protein (GFP) into the subretinal region underlying the retinas of adult mice¹. Several weeks later, the team found GFP-expressing mature photoreceptors in the retina, and concluded that the immature photoreceptors had migrated into the site and differentiated.

In a subsequent study², the same research team showed that transplantation of photoreceptor precursors could partially restore vision in mouse models of retinal degeneration. The researchers concluded that the transplanted photoreceptor precursor cells

had functionally integrated into the neural circuitry. Transplantation studies using GFP-expressing bone-marrow cells had previously revealed⁶ that donor cells could fuse with neurons, giving rise to cells with two nuclei, but the research team ruled out this possibility in the retina, showing that GFP-expressing cells in the recipient contained only a single nucleus.

Unlike the stem cells that can rejuvenate adult muscle and blood, the human retina has no stem cell that can regenerate photoreceptors⁷. However, advances in stem-cell programming have enabled the production of a renewable pool of photoreceptor precursors^{8–10}. These advances, together with the retinal-transplant research^{1,2}, set the stage for a personalized stem-cell therapy for retinal degeneration. In theory, a patient's blood or skin cells could be reprogrammed into stem cells that give rise to photoreceptor precursors for transplantation into the eye. There, they would integrate into the retina and partially restore vision.

Unfortunately, the possibility that precursors transfer molecular or genetic information to recipient photoreceptors, rather than integrating, had not been fully excluded. Material transfer would not be beneficial in advanced neurodegeneration, in which few of the patient's original photoreceptor cells remain. Singh *et al.*³, Santos-Ferreira *et al.*⁴ and Pearson *et al.*⁵ tested this possibility directly.

The three groups transplanted GFP-expressing photoreceptor precursors taken from mice up to one week old into the subretinal space of a recipient mouse retina that was labelled with the red fluorescent protein dsRED. If donor cells integrated as previously proposed, GFP-expressing cells would be surrounded by dsRED-expressing retinal cells. However, the investigators discovered that most of the GFP-expressing photoreceptors in the retina after transplantation also expressed dsRED. A series of complementary experiments ruled out the possibility of either integration or full nuclear fusion. Instead, donor cells had transferred factors (DNA, RNA or protein) that confer GFP expression to the cytoplasm of photoreceptors in the recipient retina (Fig. 1).

The 2006 experiments showed that GFP persisted in the recipient retina for up to a year¹, indicating that the transferred material is stable (although it remains to be seen whether this reflects the continuous active transfer of material or persistence after a single transfer event). Furthermore, the three current studies showed that the enzyme Cre recombinase can be transferred from donor to recipient photoreceptors, indicating that transfer is not restricted to fluorescent proteins. In addition, the restoration of vision in genetically engineered mouse models of retinal degeneration suggests that the genetic defects that cause degeneration can be at least partially rescued in a subset of recipient photoreceptors². Together,

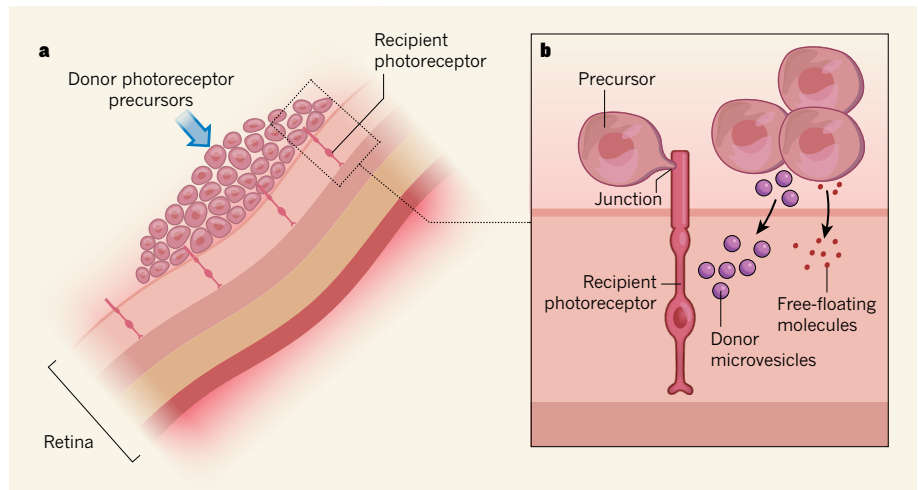


Figure 1 | Retinal transplants under scrutiny. **a**, Degeneration of photoreceptors, the retina's light-sensing cells, is associated with blindness. The injection of photoreceptor precursors from healthy mice into the subretinal region of the eyes of mice with retinal degeneration has been shown to partially restore vision. **b**, Three papers^{3–5} report that the donor precursors do not integrate into the host retina, as previously assumed, but instead transfer some of their cellular material to host photoreceptors. This could occur by means of microvesicles that bud off from the donor-cell membrane, by transfer of nucleic-acid or protein complexes through junctions between the cells, or through uptake of free-floating material across the plasma membrane.

these data point to a potential new approach to preserve photoreceptors in an injured or diseased retina — engineering cells or cellular products to efficiently transport factors required to preserve photoreceptors into the retina.

One aspect of transfer that must be carefully validated and optimized in advance of human trials is the developmental-stage specificity of material transfer from donor photoreceptor precursors. Singh *et al.* found that such precursors taken from mice between one and seven days old transfer GFP into the recipient retina more efficiently than do more highly differentiated donors. It remains to be seen whether this specificity reflects a unique

These data point to a potential new approach to preserve photoreceptors in an injured or diseased retina.

aspect of the cellular physiology of photoreceptors at this age, when they are part-way through differentiating into mature cells, or a nonspecific by-product of the process by which the cells were prepared

for transplantation. Pearson *et al.* found that GFP-expressing retinal progenitor cells in the embryo rarely transfer material to recipient photoreceptors, suggesting that only photoreceptors, and not all retinal lineages, can transfer material in this way.

Like many important advances, these three papers lead to more questions than answers. First and foremost, what are the underlying molecular and cellular mechanisms for material transfer? Perhaps transfer occurs through membrane-derived microvesicles or through protein complexes that form junctions between the cells. What is the cellular material being transferred from the donor

to the recipient — DNA, RNA or proteins? Pearson and colleagues found that subretinal injection of purified GFP protein failed to produce labelled recipient photoreceptors, suggesting that proteins are not the transferred material. However, it is possible that transfer is mediated by a larger protein complex, absent in the purified sample.

Although the findings of these studies might dampen enthusiasm for replacing photoreceptors lost during retinal degeneration in advanced cases, they also open up an exciting area of retinal research. Pursuing this avenue will advance our understanding of photoreceptors and might eventually lead to the design of methods to preserve retinal function in people with early-stage disease. ■

Michael A. Dyer is in the Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. He is also in the Department of Ophthalmology, University of Tennessee Health Sciences Center, Memphis, and the Howard Hughes Medical Institute. e-mail: michael.dyer@stjude.org

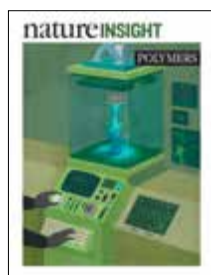
- MacLaren, R. E. *et al.* *Nature* **444**, 203–207 (2006).
- Pearson, R. A. *et al.* *Nature* **485**, 99–103 (2012).
- Singh, M. S. *et al.* *Nature Commun.* <http://dx.doi.org/10.1038/ncomms13537> (2016).
- Santos-Ferreira, T. *et al.* *Nature Commun.* **7**, 13028 (2016).
- Pearson, R. A. *et al.* *Nature Commun.* **7**, 13029 (2016).
- Weimann, J. M., Johansson, C. B., Trejo, A. & Blau, H. M. *Nature Cell Biol.* **5**, 959–966 (2003).
- Cicero, S. A. *et al.* *Proc. Natl Acad. Sci. USA* **106**, 6685–6690 (2009).
- Eiraku, M. *et al.* *Nature* **472**, 51–56 (2011).
- Meyer, J. S. *et al.* *Proc. Natl Acad. Sci. USA* **106**, 16698–16703 (2009).
- Hiller, D. *et al.* *Cell Stem Cell* **17**, 101–115 (2015).

This article was published online on 30 November 2016.

natureINSIGHT

POLYMERS





Cover illustration

Nik Spencer

Editor, *Nature*

Philip Campbell

Publishing

Richard Hughes

Insights Editor

Ursula Weiss

Production Editor

Nick Haines
Elizabeth Batty

Art Editor

Nik Spencer

Sponsorship

Reya Silao

Production

Ian Pope

Marketing

Steven Hurst

Editorial Assistant

Giacomo Russo

The Campus
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com

**SPRINGER
NATURE**

Plastic — you could be forgiven for equating it with cheap, artificial materials that have found their way into all walks of life. You might think of the frequent headlines lamenting the sheer volume of plastic waste that ends up in landfill or pollutes remote locations, such as our seas and deep ocean trenches.

But beyond all the negative headlines and commodity plastics such as packaging and plastic bags, work at the forefront of polymer research is delivering advanced materials that are helping to solve problems in areas ranging from energy and the environment to human health.

This Insight aims to provide a flavour of the opportunities offered by ‘fantastic plastic’ — polymeric materials with properties that have been precisely tailored to meet the needs of myriad low- and high-tech applications. The diversity of systems being explored and the applications being targeted are immense. This selection of reviews can cover only a fraction of them, ranging from the fundamentals of molecular design and synthesis to cutting-edge applications.

It includes a survey of the current state-of-the-art in producing more-sustainable polymers that eschew petrochemicals and use plant materials or carbon dioxide instead. There is a look at polymer-based materials that are designed to autonomously manage wear-and-tear by repairing themselves to prolong their useful lifetime, and result in regeneration and recycling after use. An examination of 3D printing using polymer-based soft materials shows how this technology is on the cusp of challenging conventional manufacturing around the world.

There is a discussion of how soft polymeric electronic materials enable devices to interface with biological tissues, facilitating new approaches to diagnostics, as well as disease prevention and control. And finally, a look at how polymer hydrogels can be crafted into objects that mimic biological structures shows how they can be put to therapeutic use.

We hope you enjoy this eclectic showcase of modern polymer research, and join us in celebrating the ingenuity of the researchers who continue to advance the field as it takes on the opportunities and challenges of the twenty-first century.

Ros Daw, Claire Hansell & Magdalena Helmer

Physical-sciences editors

CONTENTS

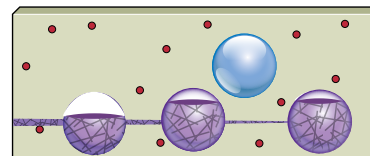
REVIEWS

354 Sustainable polymers from renewable resources

Yunqing Zhu, Charles Romain & Charlotte K. Williams

363 Polymers with autonomous life-cycle control

Jason F. Patrick, Maxwell J. Robb, Nancy R. Sottos, Jeffrey S. Moore & Scott R. White



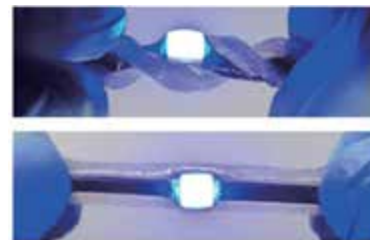
371 Printing soft matter in three dimensions

Ryan L. Truby & Jennifer A. Lewis



379 The rise of plastic bioelectronics

Takao Someya, Zhenan Bao & George G. Malliaras



386 Mimicking biological functionality with polymers for biomedical applications

Jordan J. Green & Jennifer H. Elisseeff

Sustainable polymers from renewable resources

Yunqing Zhu^{1*}, Charles Romain^{2*} & Charlotte K. Williams¹

Renewable resources are used increasingly in the production of polymers. In particular, monomers such as carbon dioxide, terpenes, vegetable oils and carbohydrates can be used as feedstocks for the manufacture of a variety of sustainable materials and products, including elastomers, plastics, hydrogels, flexible electronics, resins, engineering polymers and composites. Efficient catalysis is required to produce monomers, to facilitate selective polymerizations and to enable recycling or upcycling of waste materials. There are opportunities to use such sustainable polymers in both high-value areas and in basic applications such as packaging. Life-cycle assessment can be used to quantify the environmental benefits of sustainable polymers.

Modern life relies on polymers, from the materials that are used to make clothing, houses, cars and aeroplanes to those with sophisticated applications in medicine, diagnostics and electronics. The vast majority of these polymers are derived from petrochemicals. Many polymers contribute considerably to an improved quality of life and a cleaner environment, for example, as materials that enable the purification of water or as polymer composites with improved fuel economy for aerospace applications. Only about 6% of the oil produced worldwide is used in the manufacture of polymers, yet there are environmental concerns associated with both the raw materials used to make them¹ and their end-of-life options². Although there is no panacea for these complex environmental problems, one option is to develop more 'sustainable' polymers. Research has focused mainly on replacing fossil raw materials with renewable alternatives and on developing end-of-life options that generate materials that are suitable for recycling or biodegradation. Where biomass from plants is used as the renewable raw material, the polymers are often referred to as bioderived. In terms of biodegradation, it is important to recognize that some petrochemical polymers are biodegradable, and that not all bioderived polymers will biodegrade. The potential for sustainable polymers is stimulated by policy, legislation and international agreements, including some negotiated at the 2015 United Nations Climate Change Conference (COP21) (ref. 3) in Paris on reducing CO₂ emissions. Although the commercial application of bioderived polymers can benefit from improvements in environmental performance (as well as from supportive policy or legislation), it will also require favourable economics and material properties that are better than seen in conventional materials, including thermal resistance, mechanical strength, processability and compatibility. Taken together, these are tough criteria that could explain, in part, why there are few commercially successful sustainable polymers at present. In 2014, for example, only 1.7 megatonnes of more than 300 megatonnes of polymers produced globally were bioderived, of which the three main products, by volume, were polyethylene terephthalate (PET), polyethylene and polylactide⁴. There are two general approaches to preparing sustainable polymers: lessening the environmental impact of conventional production, for example by using biomass to make known monomers or polymers such as PET and polyethylene; and the preparation of new, 'sustainable' structures, such as polylactide, from renewable raw materials.

This Review highlights some of the opportunities for creating sustainable polymers from four renewable raw materials: carbon dioxide, terpenes, vegetable oils and carbohydrates (Fig. 1). These feedstocks enable the production of polymers and materials with a wide range of properties and applications. (The use of modified natural polymers such as cotton, silk, thermoplastic starch, cellulose derivatives and natural peptides is not discussed, although the potential for producing polymers from lignin is mentioned briefly.) The examples highlighted have been selected because they address some of the overarching challenges of sustainable polymer production.

The first challenge is that transformation of renewable resources and the production of polymers must be highly efficient to reduce costs. Production can be made more efficient by using mixtures of raw materials, producing monomers of lower purity or through the 'upcycling' of waste materials from agriculture or industry. Second, sustainable polymers must show complementary or improved properties compared with the polymers available at present. Applications with high-value markets, such as thermoplastic elastomers, rigid plastics and polyols, might present more favourable economics than applications such as packaging. Third, life-cycle assessment should be used to quantify the impact of sustainable polymers and to compare them with existing petrochemical benchmarks; this technique is usually used to assess environmental impacts and outputs that are associated with polymer production. Although the need for such comparisons might seem obvious, there are complexities associated with selecting appropriate benchmarks, boundaries and data⁵. At present, materials in the early stages of development, particularly in academic labs, are not routinely examined by life-cycle assessment. Here we highlight examples of life-cycle assessment that demonstrate the improved sustainability of the polymers, and where the findings are relevant to the design and development of future materials.

Upcycling carbon dioxide into polymers

Using waste greenhouse gases such as carbon dioxide to prepare useful and valuable polymers has long been of interest to researchers, and this chemical process is now on the cusp of commercialization. It is a rare example of a process that consumes carbon dioxide as a reagent. It enables 30–50% of a polymer's mass to be derived from carbon dioxide, with the remainder derived from petrochemicals, and it delivers both economic and environmental benefits^{6–12} (Fig. 2).

¹Chemistry Research Laboratory, Department of Chemistry, University of Oxford, Oxford OX1 3TA, UK. ²Department of Chemistry, Imperial College London, London SW7 2AZ, UK. *These authors contributed equally to this work.

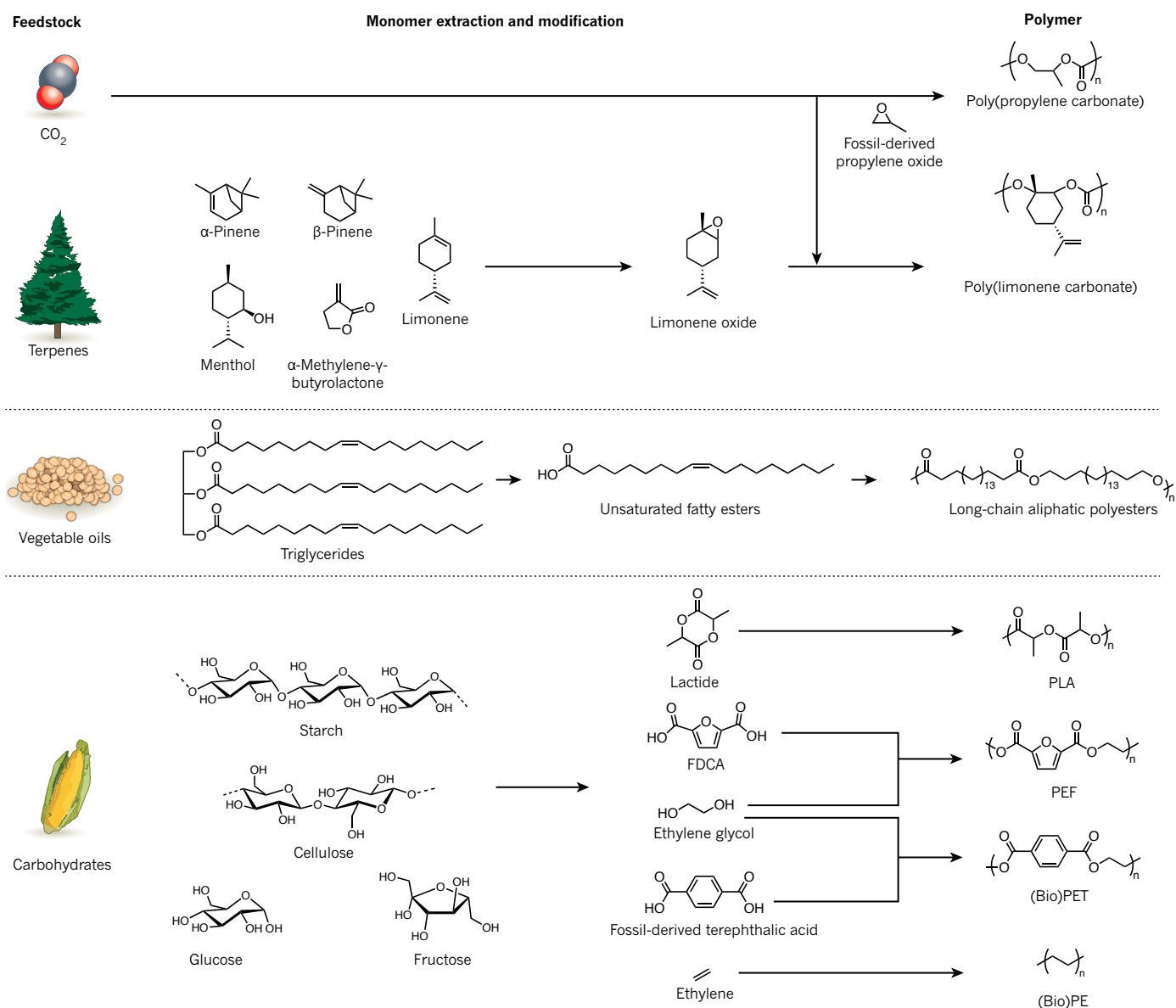


Figure 1 | Options for replacing petrochemicals as raw materials in the manufacture of polymers. Carbon dioxide is copolymerized with propylene oxide to generate propylene carbonate polyols. Terpenes, such as limonene, are chemically transformed to limonene oxide and copolymerized with carbon dioxide to generate poly(limonene carbonate). Triglycerides, from

vegetable oils, are transformed into long-chain aliphatic polyesters. Natural carbohydrate polymers, such as starch, are broken down to glucose, which is subsequently transformed to polymers such as poly(ethylene furanoate) (PEF), polylactide (PLA), bioderived poly(ethylene terephthalate) ((bio)PET) or bioderived polythene ((bio)PE).

Sustainable polymers can be produced through alternating copolymerization of epoxides, commonly propylene oxide, and carbon dioxide. The efficiency of the process is highly dependent on the catalyst that is applied, and efforts worldwide have focused on improving and understanding the underlying catalysis^{6,13–19}. Generally, homogeneous catalysts deliver a much greater uptake of carbon dioxide into the polymer, which results in balanced epoxide–carbon dioxide enchainment and produces aliphatic polycarbonates. By contrast, heterogeneous catalysts require considerably higher pressures and result in lower levels of carbon dioxide incorporation, thereby producing polyether carbonates in which the ether linkages result from sequential epoxide enchainment. It is also feasible to selectively combine the alternating copolymerization of epoxides and carbon dioxide with the polymerization of other bioderived monomers to produce block copolymers that are composed of ester, ether and carbonate blocks^{20–23}. Although these methods are at an early stage of development, they highlight the need for more-selective chemistry that uses monomer mixtures and the

potential to use block copolymers to control the macroscopic properties; both are important areas for future development.

The commercialization of polymers made with carbon dioxide addresses two distinct molecular-weight regimes and areas of application. Low-molecular-weight hydroxyl end-capped polycarbonates or polyether carbonates are applied as polyols in the manufacture of polyurethane^{6,11}. Polyols with low viscosities and low glass transition temperatures could be substitutes for some common petrochemical-based polyols that are used to make furniture foams, adhesives, clothing and resistant coatings¹⁵. Alternatively, high-molecular-weight polycarbonates are already in use as binders and sacrificial materials, which are used to ‘pattern’ a substrate before being burnt away during the fabrication process. Improvements to their properties might widen their applications to include rigid plastics and blends with petrochemical-based polymers^{12,24}.

An important benefit of upcycling carbon dioxide — although not shared by many bioderived monomers — is that polymers can be

produced easily using present infrastructure for petrochemical-based polymer manufacturing. In particular, polymerizations can proceed using existing reactors and methods for processing and purification. There is also no dependence on agriculture for raw materials or on complex pretreatments and transformations of monomers.

With a view to sustainability, an obvious question to ask is whether the manufacturing process is truly compatible with the recycling of waste carbon dioxide. The first signs are encouraging, with successful polymerization being achieved using carbon dioxide emissions that were captured from a coal-fired power station in the United Kingdom¹¹. The catalytic performance of the reaction and the quality of the product were almost equivalent to those achieved with ultrapure carbon dioxide. The entire process was surprisingly tolerant of contaminants such as water, nitrogen and oxygen, as well as any small-molecule amines and thiols that are present owing to the capture process.

Life-cycle assessment has been used to compare polyols made by the copolymerization of propylene oxide and carbon dioxide with those prepared only by propylene oxide polymerization²⁵. Even when propylene oxide was only partially substituted with carbon dioxide, the net reductions in the emission of greenhouse gases and the depletion of fossil resources were about 11–20%. It is important to emphasize that these environmental benefits arise from the replacement of the epoxide by carbon dioxide and not just from the recycling of carbon dioxide. Epoxides that are derived from limonene and vegetable oil have the potential to yield fully renewable polycarbonates^{26,27}. Poly(limonene carbonate) is qualified for use in various applications, including as a resistant and hard dry-powder coating produced by crosslinking the pendant alkene functional group that is introduced on the limonene unit^{12,26,28}. By selecting the catalyst carefully, it is also possible to prepare highly crystalline stereocomplexed poly(limonene carbonate) — a co-crystallite formed from polymer chains of opposite chirality — which has better thermal properties, including a higher degradation temperature (265 °C), than do analogues of lower crystallinity²⁹. These findings highlight the potential of bioderived materials to deliver products of high impact and value by taking advantage of naturally 'rigid' chemical functionalities and by providing cost efficiency through the use of waste as a raw material. However, there are insufficient physical, rheological and processing data to fully understand this potential.

From plants to plastics

The field of polymer science originates from studies of biopolymers such as cellulose³⁰. Many commercial sustainable polymers are sourced from plants that are rich in sugar or starch, including sugar cane (*Saccharum officinarum*), wheat (*Triticum* spp.) or sugar beet (*Beta vulgaris*). In bio-PET, for example, there is a partial substitution of a petrochemical-derived raw material: up to 30% by weight (wt%) of the ethylene glycol monomer is produced from starch. This process is complex and involves starch degradation, glucose fermentation, ethanol dehydration, ethene oxidation and hydrolysis of the product. A number of research programmes, including some already at the pilot stage, are actively investigating PET that is fully derived from biomass, in which the co-monomer terephthalic acid is also produced from

biomass³¹. Life-cycle assessment of the present generation of bio-PET shows reductions of 20–50% in the emission of greenhouse gases in comparison to petrochemical-derived PET. Like PET, polyethylene can also be produced from sugar cane, with the ethylene monomer obtained through the dehydration of ethanol. This method of polymer production is controversial and requires a well-developed sugar-cane industry; it is being explored mainly in Brazil³². After ethylene has been produced, the process of polymerization and the properties of the resulting polymer are identical to those for petrochemical-derived polyethylene. Regardless of the method of production, polyethylene usually persists in and pollutes the environment, and it is unlikely to become economically viable to recycle the material.

An advantage of developing bioderived monomers as direct substitutes — such as those used in the production of PET and polyethylene — is that the processing and applications of the resulting polymers are identical, thereby simplifying their adoption and accelerating their uptake. This is particularly important for PET, one of the few polymers for which large-scale recycling infrastructures exist, and life-cycle assessment indicates positive contributions to its sustainability if it is recycled³³. Highly efficient catalytic methods have been pioneered that enable the chemical recycling of PET and show the potential for subsequent 'upcycling' of PET into other polymers³⁴. Such methods could be a promising alternative to well-established mechanical recycling. Generally, technologies that use renewable resources to prepare polymers are front-runners in the commercialization of sustainable polymers.

It is important to consider whether using edible feedstocks to prepare polymers will have a societal impact. At first, this seems analogous to the controversy that surrounds some biofuels. However, it is apparent that polymer production is dwarfed in magnitude by that of biofuels, and bioderived materials are still a niche market in the polymer sector. A detailed study of bioderived-polymer production in the European Union substantiates the possible land-use requirements³⁵. It envisages a market share of 1–4% for bioderived polymers by the year 2020, with the exact value dependent on various economic and growth models. In a scenario in which wheat is the only source of starch, just 1–5% of the land used presently to grow wheat would be needed³⁵.

From terpenes to elastics and coatings

Terpenes and terpenoids are components of essential oils that are derived from plants and have a common isoprene unit in their chemical structures³⁶. The best-known example of a polyterpene is probably natural rubber. More than 10 megatonnes are produced per year, and the main constituent is polyisoprene. Other terpenes are being investigated as monomers for polymer production, although on a much smaller scale. These include turpentine, which is extracted from pine trees (*Pinus* spp.) and is composed mainly of α -pinene (45–97%) and β -pinene (0.5–28%), and limonene, which is extracted from the peel of citrus fruits³⁷ (Fig. 3). Worldwide production of these monomers is modest: in 2013, about 0.3 kilotonnes of turpentine³⁸ and about 0.7 kilotonnes of limonene³⁹ were produced. Commercially available polymer resins from these terpenes already exist^{36,37,39,40}.

A drawback of terpenes is the low molecular weights of their

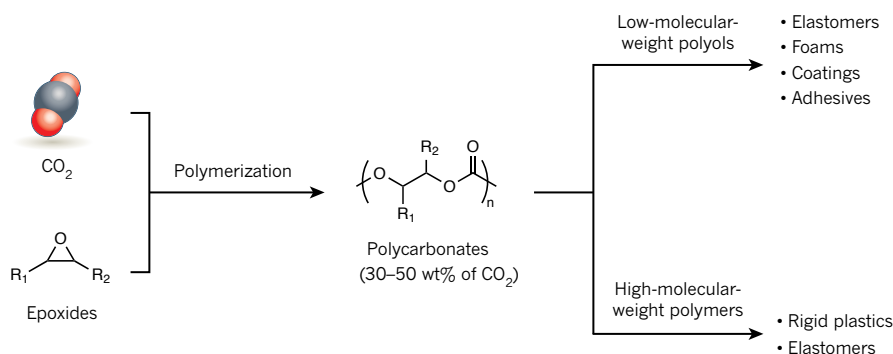


Figure 2 | Upcycling of carbon dioxide into sustainable polymers of high value.

Carbon dioxide and epoxides can be copolymerized to deliver aliphatic polycarbonates. Polycarbonate polyols of low molecular weight may be suitable to prepare foams, coatings and adhesives, whereas high-molecular-weight polycarbonates may be used as rigid plastics or elastomers.

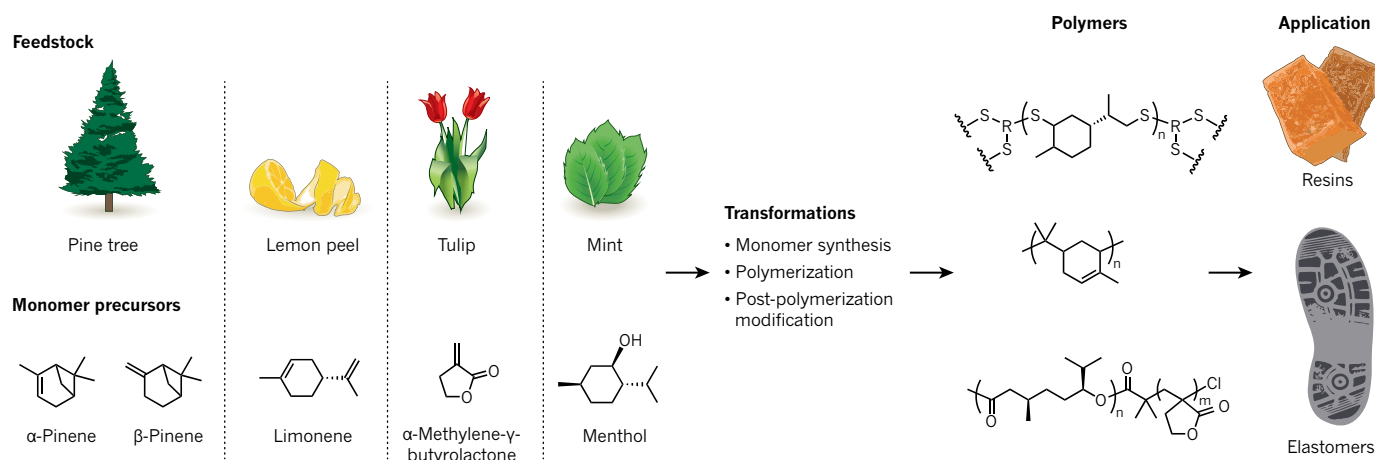


Figure 3 | Sustainable polymers produced from terpenes and terpenoids. Terpenes such as pinene and menthol are extracted from plants such as pine or mint. They can then be transformed into polymer resins or elastomers before being used.

polymers, which limits their mechanical performance. It might be possible to achieve considerably higher molecular weights through a cationic polymerization of β -pinene followed by hydrogenation⁴¹. The resulting polymer exhibits some thermal and mechanical properties that are akin to those of polymethyl methacrylate, and the material also shows high optical transmittance. Another option is to copolymerize terpenes with common petrochemical-derived vinyl monomers, such as methacrylates, by means of radical polymerization, which enables the production of materials with partial bio-based content^{40,42}. The recycling of polymers is an important aspect of sustainability. Limonene, for example, can be used both as a reagent and as a solvent for recycling⁴³. One example involves dissolving a polystyrene drinking cup in limonene; the subsequent crosslinking reaction forms an elastomer that can be moulded into a mobile-phone cover⁴³.

A further limitation in the commercialization of terpenes is their relatively high cost, which can be ameliorated through the fabrication of products of higher value such as thermoplastic elastomers (Fig. 3). Current thermoplastic elastomers are derived mostly from petrochemicals and are produced in quantities greater than 3.5 megatonnes per year for applications as diverse as car suspension systems, window seals, coatings of household goods or electronics, shoe soles or medical devices⁴⁴. The use of terpene monomers derived from wild mint (*Mentha arvensis*) and a tulip (*Tulipa gesneriana*), together with methods of controlled polymerization, has led to the production of block copolymer thermoplastic elastomers^{45,46} (Fig. 3). The bioderived polymer with the best mechanical performance has a Young's modulus of more than 6.0 megapascals, which is within the range that is observed for commercial polystyrene-butadiene-styrene (SBS). In contrast to SBS, however, the bioderived elastomer has a very high glass transition temperature (T_g) of 170–190 °C, which enables it to retain elasticity at elevated temperatures — a feature that might be desirable for applications in harsh environments. Most bio-based elastomers show elongation-at-break values of less than 1,000%. The mechanical properties remain inferior to those of petrochemical-derived polymers.

Value-added vegetable oils

Triglycerides are harvested from the seeds of certain plants, the top four of which, by volume, are soybean (*Glycine max*), oil palm (*Elaeis*), oilseed rape (*Brassica napus*) and sunflower (*Helianthus*). They are produced on a very large scale (156 megatonnes in 2012): the majority are used as food, about 30 megatonnes are used as biofuels, and about 20 megatonnes are used as chemical feedstocks⁴⁷. They also represent an important and long-standing raw material for polymer production (Fig. 4). Indeed, there is a strong track record of using linoleum (produced from linseed oil) and epoxidized oils as resins, coatings and in paints. The commercial production of polyamides from castor oil, which

is extracted from the seeds of the castor oil plant (*Ricinus communis*), yields nylon 11, nylon 6,10 and nylon 4,10. Some of these bioderived nylons have beneficial properties, including low water absorption, high chemical resistance, high temperature stability and a lack of long-term ageing⁴⁷. They have been used as toothbrush fibres, in pneumatic air-brake tubing and in flexible oil and gas pipes. An important limitation on the production of such nylons is the reliance on castor oil, which contains a secondary hydroxyl group in the fatty-acid chain that facilitates its efficient transformation to monomers and subsequent polymerizations. Notably, castor oil costs almost twice as much as more common oils such as palm oil or rapeseed oil.

Although triglycerides are found in almost all plants, the quantity that is available varies, and even common crops such as soybeans are estimated to yield only 20 wt% of triglycerides. Another challenge is that the chemical compositions of triglycerides vary both between and within a particular crop. Triglycerides are composed of three, often distinct, fatty-acid groups that are linked together through ester bonds to a glycerol unit. They are commonly processed by transesterification reactions to produce fatty esters and glycerol. In terms of polymer production, glycerol can be used as a crosslinking agent in resin production, or as a raw material for the production of monomers such as epichlorohydrin and lactic acid^{48,49}. However, the main opportunity probably comes from the fatty esters, which feature long alkyl chains (C_{12} – C_{22}) and include a considerable number of internal alkene functional groups⁵⁰. On polymerization, they can have properties that are intermediate between those of polyalkenes, such as polyethylene, and more-polar short-chain polyesters. A common set of polymerization methods makes use of the alkene groups found in unsaturated fatty esters (Fig. 4). Indeed, a considerable proportion (about 20–60 wt%) of the plant oils produced in the largest volumes worldwide is composed of such unsaturated fatty acids. More recently, crop engineering has produced a strain of soybean that yields more than 75% mono-unsaturated oleic acid — a particularly useful monomer⁵¹. Many methods exist for transforming the alkene groups to polymers, including the thiol-ene reaction, acyclic diene metathesis, epoxidation and radical or thermal crosslinking reactions^{50,52} (Fig. 4). An area that has considerable potential is reacting the alkene to produce α,ω -diesters or α,ω -diols. These monomers undergo conventional condensation polymerizations to yield bioderived polyesters, and if α,ω -diamides are used as the monomer, nylons can be produced. One limitation is that α,ω -difunctionalized monomers are usually produced in reactions such as olefin metathesis, ozonolysis or oxidative cleavage of the carbon–carbon double bond, and in all cases, only about half of the fatty acids are used and several by-products are produced⁵³. An elegant solution involves the use of selective chemical catalysis to isomerize the internal alkene group to the chain end. Following an alkoxycarbonylation process, this enables near-quantitative

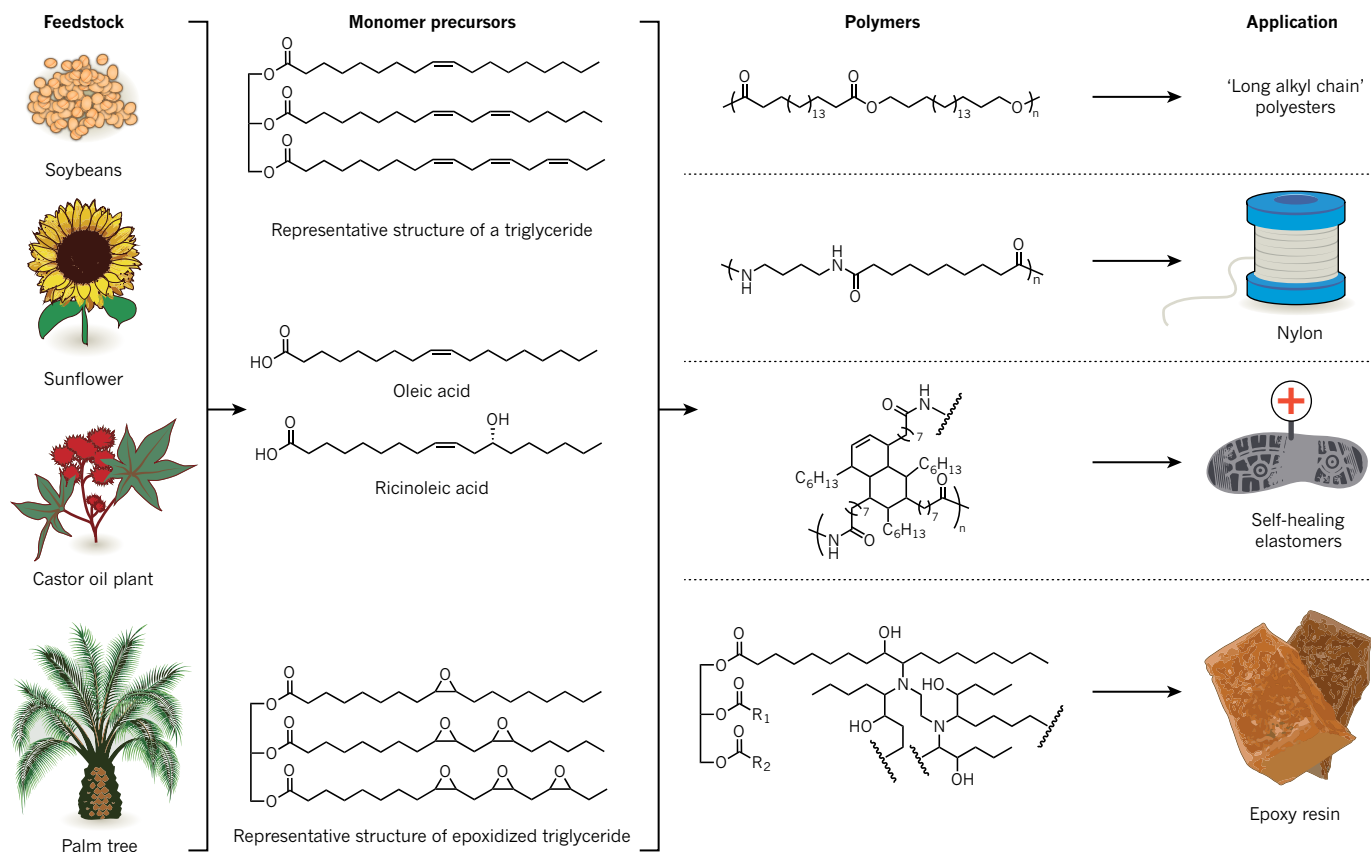


Figure 4 | Sustainable polymers produced from vegetable oils. Plants such as soybean, sunflower, castor oil or palm tree are good sources of triglycerides. The triglycerides are transformed to polymers such as polyesters or nylons and are subsequently applied as elastomers or resins.

production of the desired α,ω -difunctionalized monomers^{53–55}. The diesters are produced with more than 95% selectivity, at high conversion rates and without significant by-products^{53–55}. Polyesters can be prepared through conventional polycondensations and the resulting materials have thermal properties, solid-state crystalline structures and tensile properties similar to those of polyethylene⁵⁶. Enzyme catalysts that enable similar polycondensations have also been developed that show efficient activity over a range of temperatures, solvents and substrates⁵⁷. Despite the ability to process the long-chain polyester products using existing industrial methods such as injection moulding or film extrusion⁵⁴, these bioderived polyesters have so far been unable to compete with the cost of petrochemical-derived polyethylene. However, the high crystallinity, thermal and chemical resistance and degradability of bioderived polyesters are still valuable properties, and the application of these materials as compatibilizers in blends of petrochemicals or as macromonomers is feasible.

Polycondensation requires monomers of very high purity and balanced stoichiometry to successfully produce polymers. An alternative method applies to macrolactones, which are derived from fatty acids and can undergo ring-opening polymerization to produce similar long-alkyl-chain polyesters. It enables the production of high-molecular-weight polymers and is compatible with block copolymerization. Bioderived macrolactones that consist of up to 23 atoms have been polymerized⁵⁸, and some macrolactones, such as pentadecalactone or ambrettolide, are naturally occurring biochemicals. The resulting polyesters have thermal and rheological properties akin to those of linear low-density polyethylene⁵⁹. In a biocatalytic approach, mixtures of glucose and oleic acid were fermented to efficiently produce more than 200 grams per litre of the macrolactone lactonic sophorolipid, which features both disaccharide and alkene functional groups. The ring-opening metathesis polymerization of this macrolactone leads to the production of carbohydrate functionalized polyesters⁶⁰.

Concerns regarding land use can arise because many triglycerides are also used as foods. One solution might be to engineer algae to biosynthesize unsaturated fatty acids. Algae can be grown on non-arable land and might even flourish in brackish water. They also require only sunlight and carbon dioxide as sources of energy. Algal biosynthesis of triglycerides has enabled dry-weight yields of 20–50%, which is higher than the yields of many crops⁶¹.

Triglycerides have also been used to prepare thermoplastic elastomers, including a self-healing and thermoreversible elastomer⁶². The Young's modulus of this material is comparable to that of the petrochemical-derived polymer SBS, and its maximum strain exceeds 500%. The use of reversible supramolecular hydrogen-bonding interactions for crosslinking facilitates its processing, and the elastomer might even be recyclable — which is not usually possible for conventional elastomers. Although mechanical creep and other rheological properties have not yet been reported, the diversity of available fatty acids and the novelty of the methods for tuning the physical properties are attractive. Fatty acids can also form vitrimers, which are polymers that show reversible temperature-induced thermoset to thermoplastic transitions, for example through thermally controllable transesterification reactions^{63,64}.

Sugars as sustainable materials

Each year, more than 150 billion tonnes of polysaccharides are produced naturally, with humans consuming only about 1% of this volume. To make synthetic polymers, these biopolymers must be separated and depolymerized to obtain monosaccharides known as pentoses and hexoses. The most abundant is glucose: at present, glucose is obtained through the saccharification of starch or sucrose hydrolysis, but in the future it could come from lignocellulosic sources. Glucose is transformed into building-block chemicals such as lactic acid or succinic acid, which are polymerized directly or are reacted further through chemical or enzymatic routes to produce monomers^{65,66} (Fig. 5).

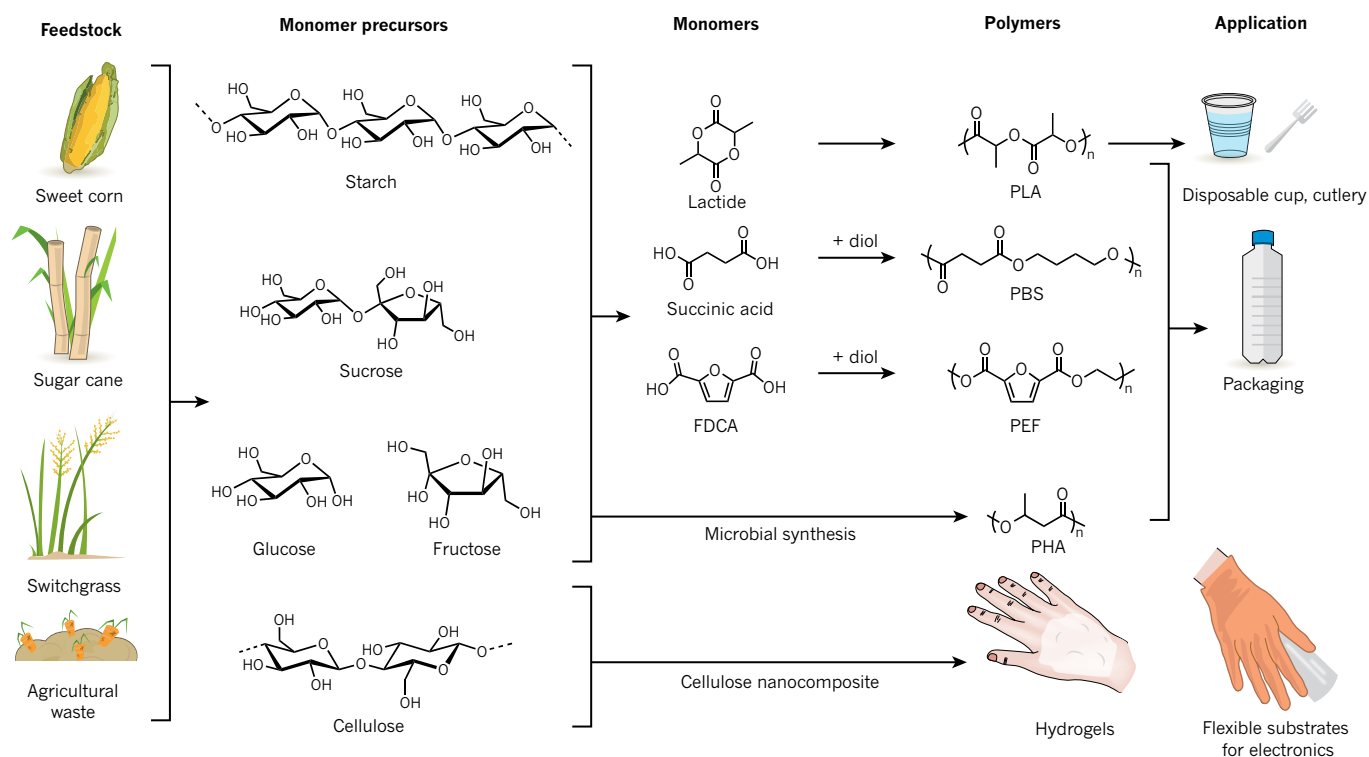


Figure 5 | Sustainable polymers produced from polysaccharides. Plants such as sugar cane and maize are good sources of sucrose or starch, which can be transformed to monomers, including lactide, succinic acid and 2,5-furandicarboxylic acid (FDCA). The monomers are polymerized to

produce polylactide (PLA), poly(butylene succinate) or poly(ethylene furanoate) (PEF), respectively. Poly(hydroxyalkanoate) (PHA) may be produced directly from glucose by biosynthesis. Cellulose fibres can be used to reinforce composites for use as hydrogels or flexible substrates for electronics.

In 2004, the US Department of Energy published a landmark report that highlighted the top value-added chemicals from biomass, which were selected on the basis of their availability and the ability to transform them into useful products⁶⁷. These molecules include lactic acid, succinic acid and 2,5-furandicarboxylic acid (FDCA), which have already delivered useful polymers (Fig. 5). In some cases, it is necessary to use new polymerization methods and processes because the monomers involved are more polar and highly functionalized (that is, oxygenated) than are those derived from fossil raw materials.

One important example is commercially available polylactide (PLA), which is made from starch-rich crops such as maize (corn; *Zea mays*) (Fig. 5). PLA is produced through the fermentation of starch to lactic acid, followed by the preparation of lactide and its subsequent polymerization. It has properties that enable it to replace petrochemical-derived plastics in some types of packaging and fibres^{68,69}. The fabrication process requires the efficient production of lactic acid and lactide, the former of which is achieved through microbial fermentation^{70,71}. Advances, including the use of cheaper fermentation substrates such as glycerol, agricultural waste and even algae-produced carbohydrates⁷⁰, may improve efficiency and profitability. Chemical catalysis might also be of interest as a means to produce racemic mixtures of lactide from sugars⁷². The selective polymerization of such mixtures might increase the thermal resistance and range of applications of PLA⁷³. In terms of its end-of-life fate, PLA can be recycled and degraded. It is even compostable at high temperatures, degrading to lactic acid, which can be metabolized naturally. Life-cycle assessment has shown reductions of up to 40% in greenhouse-gas emissions and up to 25% in non-renewable-energy use for PLA compared with petrochemical-derived polymers such as polyethylene or PET^{74,75}. However, the production of PLA might have other environmental impacts, including the use of water and fertilizer, which are more difficult to compare with the impacts of fossil-fuel extraction, purification and storage. Another substantial hurdle is to replace virgin crops (such as maize or sugar cane) with lignocellulosic or waste biomass⁷⁶.

Another group of renewably sourced polyesters are the polyhydroxyalkanoates (PHAs), which are obtained through the fermentation of sugar⁷⁷ (Fig. 5). They occur naturally and can be harvested in excellent yields from microorganisms directly, without the need for intermediate monomer isolation. Biosynthesis is achieved by culturing bacteria under growth-limiting conditions and results in the accumulation of considerable quantities of the polymer in the cytoplasm. The most promising PHAs have similar physical properties to polyalkenes such as polypropylene, but also offer the advantage of degradability. Production in bacteria is not cost-effective for commodity applications at present, although small-scale production is being explored and might be suitable for higher-value medical applications⁶⁵.

Polyethylene furanoate (PEF) is an attractive example of a fully bio-derived material with properties that make it suitable as a substitute for PET in some applications. Although it has not yet been commercialized, pilot-scale production of PEF seems to be under way. It is produced by the transformation of fructose or glucose to hydroxymethyl furfural (HMF) through acidification and dehydration reactions⁷⁸. HMF is unstable, which limits the efficiency of the process and results in side products such as levulinic acid. An improved route to HMF ethers, which are more stable and can be oxidized to FDCA, has been reported⁷⁹ (Fig. 5). FDCA is then copolymerized by polycondensation with bio-derived ethylene glycol to yield fully bio-derived PEF. Importantly, both the polymerization and oxidation reactions are compatible with PET manufacturing, and this potential to use existing infrastructure might accelerate the translation and uptake of bio-derived PEF⁷⁹. PEF has a higher T_g and improved barrier properties, especially with respect to oxygen permeability, than does PET⁷⁹, and it is less likely to undergo cold crystallization. A life-cycle assessment that benchmarked bio-derived PEF against petrochemical-derived PET showed a reduction in greenhouse-gas emissions of up to 55% (refs 78 and 80). It is difficult to compare the costs of the two materials because of the disparity in scales of production; however, larger-scale production of PEF will probably reduce its cost⁷⁸.

Succinic acid is an important monomer that is derived from the highly efficient fermentation of glucose and is produced in quantities of 170 kilotonnes per year⁸¹. It can be reacted through polycondensation with bioderived 1,4-butanediol to produce PBS, which is produced commercially in quantities of about 40 kilotonnes per year. Polybutylene succinate is a semicrystalline polymer with a high melting temperature (T_m) of 115 °C, and it can be processed using some conventional techniques, although it presents rheological limitations to the production of blown films⁸¹. It has been used as a barrier in packaging and also in blends. An alternative method for producing related polyesters with repeated succinic acid units is to copolymerize epoxides with succinic anhydride⁸². The method is attractive because it can be controlled; it might obviate the need for precise control of reagent stoichiometry and it yields materials with predictable molecular weights. Stereocomplexes of polypropylene succinate have been created that are crystalline and thermally resistant materials with T_m values of around 120 °C (ref. 83). Succinic acid can also be dehydrated at elevated temperatures to produce γ -butyrolactone. Historically, it was considered to be impossible to polymerize this five-membered ring lactone owing to its low ring-strain. However, optimized low-temperature processes involving *in situ* polymer precipitation have enabled the production of some polymer⁸⁴. Despite issues that impede the commercial deployment of polymerized γ -butyrolactone, including the method of production, the stability of the polymer and that the monomer is a controlled substance⁸⁵, this example demonstrates the potential for selective catalysis and the recycling of bioderived monomers and polymers.

Carbohydrates could provide a more cost-effective route to thermoplastic elastomers than terpenes. In particular, they show potential as block copolyester elastomers. Engineered *Escherichia coli* bacteria have been used to prepare a functionalized lactone at high efficiency (88 grams per litre in semi-batch mode), with the cost of the monomer estimated at US\$2 per kilogram, which is within the acceptable range for some commodity applications⁸⁶. The lactone is polymerized using controlled ring-opening polymerization to produce an elastomer with a T_g of -50 °C, and its copolymerization with polylactide yields a thermoplastic elastomer that can be stretched to 18 times its original length without breaking.

It has been known for more than a century that cellulose can be used to produce commercial polymers such as cellophane or cellulose acetate. Cellulose fibres are also used as reinforcements in natural-fibre-polymer composites, which makes them attractive as engineering materials^{87–89}. Semi-renewable hydrogels have been prepared by polymerization from the hydroxyl groups of hemicellulose that is harvested from the Norway spruce (*Picea abies*)⁹⁰. The method is straightforward, tolerant of the reaction conditions, and allows control of the crosslinking density, which enables tailoring of the polymer's ultimate mechanical properties. In an alternative approach, cellulose nanofibrils derived from wood have been used to replace PET as the flexible substrate in electronics manufacture. The cellulose fibrils show a high electrical breakdown tolerance (up to 1,100 volts), and the paper product undergoes fungal biodegradation without adverse environmental effects⁹¹.

Outlook and future prospects

Sustainable polymers from renewable resources are already gaining importance, and in the future, society will both want and need materials that have a smaller ecological footprint. Early successes in creating sustainable polymers have led to commercial products that are mostly used in packaging and as fibres. An important challenge is to identify platform chemicals or building blocks that can be easily prepared from abundant feedstocks and that do not compete for resources with food crops or alter the ecosystem.

Improvements to agricultural methods for crop production and harvesting, for example to optimize yields, are likely both to enhance the economic impact and to ease the environmental impact of bioderived polymers. Research should include ways to make better use of waste from agriculture and industry as monomers, including corn stover, fruit

pulp, forestry waste and carbon dioxide emissions. Another important opportunity arises from the ready availability worldwide of lignocellulosic biomass: however, improved biopolymer separation, degradation and transformation chemistry and biochemistry will be required to optimize both the yield and cost of the monomers. Carbohydrates are the most abundant and easily processed sources of sustainable monomers, and a number of interesting carbohydrate-based processes and polymers are already being developed. To prepare high-value products, research needs to exploit the high degree of natural functionality of carbohydrates, which includes taking advantage of rigid carbohydrate ring structures, as well as the extensive opportunities for non-covalent interactions and stereoregularity that carbohydrates offer^{92,93}. In tandem, the transformation of lignins to polymers has been underdeveloped owing to the highly complex and changeable structures of the lignins⁹⁴. Further methods to selectively transform lignin into monomers are needed: studies have highlighted the potential for catalysis to deliver these monomers, although so far they have focused on model compounds rather than native lignin^{95,96}. Another interesting option is to apply ferulic acid, which can be derived from either lignin or agricultural waste, to yield interesting polyesters and resins that are suitable as substitutes for petrochemical-derived polymers^{97,98}. A crucial, and sometimes underestimated, design criterion is the need to prioritize routes to monomers and polymers that are compatible with existing industrial infrastructure. The low cost and efficient purification of raw materials and products also requires much greater optimization for the highly oxygenated bio-based materials. The ability to retrofit existing manufacturing infrastructure to enable sustainable polymer production will continue to be an important driver in reducing costs and accelerating implementation.

The continued use of sustainable polymers in disposable applications such as packaging will result in the end-of-life fate exerting a considerable influence on sustainability. Innovative recycling, degradation or disposal options are likely to become even more important for preventing new materials from contributing to existing plastic waste issues, and there might also be an opportunity for supporting policy and legislation to shape the outcome. Although the direct quantification and comparison of sustainable polymers with petrochemical-derived equivalents is at an early stage, there have been sufficient studies to demonstrate that in many cases, the impacts of production are reduced, particularly on greenhouse-gas emissions and the depletion of fossil resources. Studies should also consider the life of the product beyond manufacture and the impacts associated with disposal. So far, few polymers have been designed to be both fully bioderived and degradable, although aliphatic polyesters such as polylactide are notable successes.

Packaging is an important opportunity at present for the application of bio-based polymers; however, it is challenging for such materials to compete economically with petrochemical-derived polymers. Instead, bio-based polymers should seek to compete in the higher-value and higher-performance application areas, including thermoplastic elastomers, engineering plastics or composite materials. To facilitate success, it will be important to tailor and improve the properties of such polymers. For example, the preparation of polymers with higher thermal resistance would enable them to compete with existing semi-aromatic polyesters and nylons. And elastomers that show greater elongation-at-break values would be able to compete with petrochemical-derived polymers. Understanding and engineering the degradation profiles of bio-based polymers, for example by combining long-term durability with triggered degradation, represents a further challenge for research.

The task of widening the scope and range for sustainable polymers is considerable; to solve these complex problems, researchers will need to work together across the conventional disciplines of agriculture, biology, biochemistry, catalysis, polymer chemistry, materials science, engineering, environmental impact assessment, economics and policy. In the future, society will need more materials that have been made efficiently from natural waste and that are suitable for recycling or biodegradation. ■

Received 20 October 2015; accepted 29 June 2016.

1. van der Ploeg, F. Natural resources: curse or blessing? *J. Econ. Lit.* **49**, 366–420 (2011).
2. Jambeck, J. R. *et al.* Plastic waste inputs from land into the ocean. *Science* **347**, 768–771 (2015).
3. Philp, J. C., Bartsev, A., Ritchie, R. J., Baucher, M.-A. & Guy, K. Bioplastics science from a policy vantage point. *New Biotechnol.* **30**, 635–646 (2013).
4. Shen, L., Worrell, E. & Patel, M. Present and future development in plastics from biomass. *Biofuel. Bioprod. Bior.* **4**, 25–40 (2010).
5. Talon, O. in *Environmental Impact of Polymers* (eds Hamaide, T., Deterre, R., & Feller, J.-F.) Ch. 6, 91–107 (John Wiley, 2014).
6. Lee, S. H., Cyriac, A., Jeon, J. Y. & Lee, B. Y. Preparation of thermoplastic polyurethanes using *in situ* generated poly(propylene carbonate)-diols. *Polym. Chem.* **3**, 1215–1220 (2012).
This paper demonstrates the use of polycarbonate polyols produced using carbon dioxide as a monomer to make polyurethanes.
7. von der Assen, N., Voll, P., Peters, M. & Bardow, A. Life cycle assessment of CO₂ capture and utilization: a tutorial review. *Chem. Soc. Rev.* **43**, 7982–7994 (2014).
8. Markewitz, P. *et al.* Worldwide innovations in the development of carbon capture technologies and the utilization of CO₂. *Energy Environ. Sci.* **5**, 7281–7305 (2012).
9. Ren, W. M., Liu, Z. W., Wen, Y. Q., Zhang, R. & Lu, X. B. Mechanistic aspects of the copolymerization of CO₂ with epoxides using a thermally stable single-site cobalt(III) catalyst. *J. Am. Chem. Soc.* **131**, 11509–11518 (2009).
10. Ellis, W. C. *et al.* Copolymerization of CO₂ and meso epoxides using enantioselective β -diiminate catalysts: a route to highly isotactic polycarbonates. *Chem. Sci.* **5**, 4004–4011 (2014).
11. Chapman, A. M., Keyworth, C., Kember, M. R., Lennox, A. J. J. & Williams, C. K. Adding value to power station captured CO₂: tolerant Zn and Mg homogeneous catalysts for polycarbonate polyol production. *ACS Catal.* **5**, 1581–1588 (2015).
This paper highlights the use of carbon dioxide captured from a power station in the United Kingdom in the production of polycarbonate polyols.
12. Hauenstein, O., Reiter, M., Agarwal, S., Rieger, B. & Greiner, A. Bio-based polycarbonate from limonene oxide and CO₂ with high molecular weight, excellent thermal resistance, hardness and transparency. *Green Chem.* **18**, 760–770 (2016).
This paper demonstrates the production and scale-up of fully bio-based polylimonene carbonate, prepared through the copolymerization of carbon dioxide with epoxide, and evaluates its properties.
13. Inoue, S., Koinuma, H. & Tsuruta, T. Copolymerization of carbon dioxide and epoxide. *J. Polym. Sci. B* **7**, 287–292 (1969).
14. Paul, S. *et al.* Ring-opening copolymerization (ROCOP): synthesis and properties of polyesters and polycarbonates. *Chem. Commun.* **51**, 6459–6479 (2015).
15. Darensbourg, D. J. Making plastics from carbon dioxide: salen metal complexes as catalysts for the production of polycarbonates from epoxides and CO₂. *Chem. Rev.* **107**, 2388–2410 (2007).
16. Lu, X. B., Ren, W. M. & Wu, G. P. CO₂ copolymers from epoxides: catalyst activity, product selectivity, and stereochemistry control. *Acc. Chem. Res.* **45**, 1721–1735 (2012).
17. Klaus, S., Lehenmeier, M. W., Anderson, C. E. & Rieger, B. Recent advances in CO₂/epoxide copolymerization — new strategies and cooperative mechanisms. *Coordin. Chem. Rev.* **255**, 1460–1479 (2011).
18. Coates, G. W. & Moore, D. R. Discrete metal-based catalysts for the copolymerization of CO₂ and epoxides: discovery, reactivity, optimization, and mechanism. *Angew. Chem. Int. Edn Engl.* **43**, 6618–6639 (2004).
19. Nozaki, K., Nakano, K. & Hiyama, T. Optically active polycarbonates: asymmetric alternating copolymerization of cyclohexene oxide and carbon dioxide. *J. Am. Chem. Soc.* **121**, 11008–11009 (1999).
20. Darensbourg, D. J. & Wu, G. P. A. One-pot synthesis of a triblock copolymer from propylene oxide/carbon dioxide and lactide: intermediacy of polyol initiators. *Angew. Chem. Int. Edn Engl.* **52**, 10602–10606 (2013).
21. Jeske, R. C., Rowley, J. M. & Coates, G. W. Pre-rate-determining selectivity in the terpolymerization of epoxides, cyclic anhydrides, and CO₂: a one-step route to diblock copolymers. *Angew. Chem. Int. Edn Engl.* **47**, 6041–6044 (2008).
22. Jeon, J. Y., Eo, S. C., Varghese, J. K. & Lee, B. Y. Copolymerization and terpolymerization of carbon dioxide/propylene oxide/phthalic anhydride using a (salen)Co(III) complex tethering four quaternary ammonium salts. *Beilstein J. Org. Chem.* **10**, 1787–1795 (2014).
23. Zhu, Y., Romain, C. & Williams, C. K. Selective polymerization catalysis: controlling the metal chain end group to prepare block copolyesters. *J. Am. Chem. Soc.* **137**, 12179–12182 (2015).
24. Luinstra, G. A. Poly(propylene carbonate), old copolymers of propylene oxide and carbon dioxide with new interests: catalysis and material properties. *Pol. Rev.* **48**, 192–219 (2008).
25. von der Assen, N. & Bardow, A. Life cycle assessment of polyols for polyurethane production using CO₂ as feedstock: insights from an industrial case study. *Green Chem.* **16**, 3272–3280 (2014).
This paper presents a life-cycle assessment that compares the production of polyols for polyurethane manufacture from petrochemical sources or through partial substitution with carbon dioxide.
26. Byrne, C. M., Allen, S. D., Lobkovsky, E. B. & Coates, G. W. Alternating copolymerization of limonene oxide and carbon dioxide. *J. Am. Chem. Soc.* **126**, 11404–11405 (2004).
27. Winkler, M., Romain, C., Meier, M. A. R. & Williams, C. K. Renewable polycarbonates and polyesters from 1,4-cyclohexadiene. *Green Chem.* **17**, 300–306 (2015).
28. Li, C., Sablong, R. J. & Koning, C. E. Synthesis and characterization of fully-bio-based α,ω -dihydroxyl poly(limonene carbonate)s and their initial evaluation in coating applications. *Eur. Polym. J.* **67**, 449–458 (2015).
29. Auriemma, F. *et al.* Stereocomplexed poly(limonene carbonate): a unique example of the cocrystallization of amorphous enantiomeric polymers. *Angew. Chem. Int. Edn Engl.* **54**, 1215–1218 (2015).
This paper demonstrates the production of polylimonene carbonate stereocomplexes through efficient catalysis.
30. Klemm, D., Heublein, B., Fink, H. P. & Bohn, A. Cellulose: fascinating biopolymer and sustainable raw material. *Angew. Chem. Int. Edn Engl.* **44**, 3358–3393 (2005).
31. Pang, J. *et al.* Synthesis of ethylene glycol and terephthalic acid from biomass for producing PET. *Green Chem.* **18**, 342–359 (2016).
32. Zhang, M. & Yu, Y. Dehydration of ethanol to ethylene. *Ind. Eng. Chem. Res.* **52**, 9505–9514 (2013).
33. Welle, F. Twenty years of PET bottle to bottle recycling — an overview. *Resour. Conserv. Recycling* **55**, 865–875 (2011).
34. Fukushima, K. *et al.* Organocatalytic depolymerization of poly(ethylene terephthalate). *J. Polym. Sci. A* **49**, 1273–1281 (2011).
35. Crank, M. *et al.* Techno-economic Feasibility of Large-scale Production of Bio-based Polymers in Europe. Technical Report EUR 22103 EN (European Communities, 2005).
36. Winnacker, M. & Rieger, B. Recent progress in sustainable polymers obtained from cyclic terpenes: synthesis, properties, and application potential. *ChemSusChem* **8**, 2455–2471 (2015).
37. Gandini, A. & Lacerda, T. M. From monomers to polymers from renewable resources: recent advances. *Prog. Polym. Sci.* **48**, 1–39 (2015).
38. Gandini, A. The irruption of polymers from renewable resources on the scene of macromolecular science and technology. *Green Chem.* **13**, 1061–1083 (2011).
39. Ciriminna, R., Lomeli-Rodriguez, M., Demma Cara, P., Lopez-Sanchez, J. A. & Pagliaro, M. Limonene: a versatile chemical of the bioeconomy. *Chem. Commun.* **50**, 15288–15296 (2014).
40. Wilbon, P. A., Chu, F. & Tang, C. Progress in renewable polymers from natural terpenes, terpenoids, and rosin. *Macromol. Rapid Commun.* **34**, 8–37 (2013).
41. Satoh, K. *et al.* Sustainable cycloolefin polymer from pine tree oil for optoelectronics material: living cationic polymerization of β -pinene and catalytic hydrogenation of high-molecular-weight hydrogenated poly(β -pinene). *Polym. Chem.* **5**, 3222–3230 (2014).
42. Sharma, S. & Srivastava, A. K. Alternating copolymers of limonene with methyl methacrylate: kinetics and mechanism. *J. Macromol. Sci. A* **40**, 593–603 (2003).
43. Hearon, K. *et al.* A high-performance recycling solution for polystyrene achieved by the synthesis of renewable poly(thioether) networks derived from D-limonene. *Adv. Mater.* **26**, 1552–1558 (2014).
44. Albertsson, A.-C., Voepel, J., Edlund, U., Dahlman, O. & Soderqvist-Lindblad, M. Design of renewable hydrogel release systems from fiberboard mill wastewater. *Biomacromolecules* **11**, 1406–1411 (2010).
45. Shin, J., Lee, Y., Tolman, W. B. & Hillmyer, M. A. Thermoplastic elastomers derived from menthene and tulipalin A. *Biomacromolecules* **13**, 3833–3840 (2012).
This paper describes how a monomer derived from wild mint (*Mentha arvensis*) can be copolymerized with one from a tulip (*Tulipa gesneriana*) to produce fully bio-based block copolyester thermoplastic elastomers.
46. Bolton, J. M., Hillmyer, M. A. & Hoyer, T. R. Sustainable thermoplastic elastomers from terpene-derived monomers. *ACS Macro Lett.* **3**, 717–720 (2014).
47. Stempfle, F., Ortmann, P. & Mecking, S. Long-chain aliphatic polymers to bridge the gap between semicrystalline polyolefins and traditional polycondensates. *Chem. Rev.* **116**, 4597–4641 (2016).
48. Santacesaria, E. *et al.* Chemical and technical aspects of the synthesis of chlorohydrins from glycerol. *Ind. Eng. Chem. Res.* **53**, 8939–8962 (2014).
49. Sharninghausen, L. S., Campos, J., Manas, M. G. & Crabtree, R. H. Efficient selective and atom economic catalytic conversion of glycerol to lactic acid. *Nature Commun.* **5**, 5084 (2014).
50. Maisonneuve, L., Lebarbe, T., Grau, E. & Cramail, H. Structure–properties relationship of fatty acid-based thermoplastics as synthetic polymer mimics. *Polym. Chem.* **4**, 5472–5517 (2013).
51. De Maria, G. Plenish™ high oleic soybean oil. The first biotech soybean product with consumer nutrition benefits. *Agro Food Ind. High-Tech* **24**, 10–11 (2013).
52. Meier, M. A. R., Metzger, J. O. & Schubert, U. S. Plant oil renewable resources as green alternatives in polymer science. *Chem. Soc. Rev.* **36**, 1788–1802 (2007).
53. Goldbach, V., Roesle, P. & Mecking, S. Catalytic isomerizing ω -functionalization of fatty acids. *ACS Catal.* **5**, 5951–5972 (2015).
54. Stempfle, F., Ritter, B. S., Mulhaupt, R. & Mecking, S. Long-chain aliphatic polyesters from plant oils for injection molding, film extrusion and electrospinning. *Green Chem.* **16**, 2008–2014 (2014).
This paper reveals how plant oils can be converted by means of highly selective catalysis to produce polyesters with properties that mimic polyethylene.
55. Witt, T., Stempfle, F., Roesle, P., Häußler, M. & Mecking, S. Unsymmetrical α,ω -difunctionalized long-chain compounds via full molecular incorporation of fatty acids. *ACS Catal.* **5**, 4519–4529 (2015).
56. Liu, C. *et al.* Polymers from fatty acids: poly(ω -hydroxyl tetradecanoic acid) synthesis and physico-mechanical studies. *Biomacromolecules* **12**, 3291–3298 (2011).

57. Gross, R. A., Ganesh, M. & Lu, W. Enzyme-catalysis breathes new life into polyester condensation polymerizations. *Trends Biotechnol.* **28**, 435–443 (2010).
58. Witt, T. & Mecking, S. Large-ring lactones from plant oils. *Green Chem.* **15**, 2361–2364 (2013).
59. Pepels, M. P. F., Koeken, R. A. C., van der Linden, S. J. J., Heise, A. & Duchateau, R. Mimicking (linear) low-density polyethylenes using modified polymacrolactones. *Macromolecules* **48**, 4779–4792 (2015).
60. Peng, Y., Decatur, J., Meier, M. A. R. & Gross, R. A. Ring-opening metathesis polymerization of a naturally derived macrocyclic glycolipid. *Macromolecules* **46**, 3293–3300 (2013).
61. Roesle, P. *et al.* Synthetic polyester from algae oil. *Angew. Chem. Int. Edn Engl.* **53**, 6800–6804 (2014).
62. Cordier, P., Tournilhac, F., Soulie-Ziakovic, C. & Leibler, L. Self-healing and thermoreversible rubber from supramolecular assembly. *Nature* **451**, 977–980 (2008).
63. Montarnal, D., Capelot, M., Tournilhac, F. & Leibler, L. Silica-like malleable materials from permanent organic networks. *Science* **334**, 965–968 (2011).
64. Altuna, F. I., Pettarin, V. & Williams, R. J. J. Self-healable polymer networks based on the cross-linking of epoxidised soybean oil by an aqueous citric acid solution. *Green Chem.* **15**, 3360–3366 (2013).
65. Chen, G. Q. & Patel, M. K. Plastics derived from biological sources: present and future: a technical and environmental review. *Chem. Rev.* **112**, 2082–2099 (2012).
- This review provides a techno–environmental assessment of bio-based polymers and monomers.**
66. Galbis, J. A., García-Martín, M. G., de Paz, M. V. & Galbis, E. Synthetic polymers from sugar-based monomers. *Chem. Rev.* **116**, 1600–1636 (2016).
67. Bozell, J. J. & Petersen, G. R. Technology development for the production of bio-based products from biorefinery carbohydrates—the US Department of Energy’s “Top 10” revisited. *Green Chem.* **12**, 539–554 (2010).
68. Auras, R., Harte, B. & Selke, S. An overview of polylactides as packaging materials. *Macromol. Biosci.* **4**, 835–864 (2004).
69. Inkinen, S., Hakkarainen, M., Albertsson, A. C. & Sodergard, A. From lactic acid to poly(lactic acid) (PLA): characterization and analysis of PLA and its precursors. *Biomacromolecules* **12**, 523–532 (2011).
70. Abdel-Rahman, M. A., Tashiro, Y. & Sonomoto, K. Recent advances in lactic acid production by microbial fermentation processes. *Biotechnol. Adv.* **31**, 877–902 (2013).
71. Dusselier, M., Van Wouwe, P., Dewaele, A., Jacobs, P. A. & Sels, B. F. Shape-selective zeolite catalysis for bioplastics production. *Science* **349**, 78–80 (2015).
72. Dusselier, M., Van Wouwe, P., Dewaele, A., Makshina, E. & Sels, B. F. Lactic acid as a platform chemical in the biobased economy: the role of chemocatalysis. *Energy Environ. Sci.* **6**, 1415–1442 (2013).
73. Ikada, Y., Jamshidi, K., Tsuji, H. & Hyon, S. H. Stereocomplex formation between enantiomeric poly(lactides). *Macromolecules* **20**, 904–906 (1987).
74. Shen, L., Worrell, E. & Patel, M. K. Comparing life cycle energy and GHG emissions of bio-based PET, recycled PET, PLA, and man-made cellulotics. *Biofuel. Bioprod. Bior.* **6**, 625–639 (2012).
75. Groot, W. J. & Borén, T. Life cycle assessment of the manufacture of lactide and PLA biopolymers from sugarcane in Thailand. *Int. J. Life Cycle Assess.* **15**, 970–984 (2010).
76. Corbion. Corbion Purac successfully develops PLA resin from second generation feedstocks. *Corbion* <http://www.corbion.com/media/press-releases?newsId=1955535> (2015).
77. Müller, H.-M. & Seebach, D. Poly(hydroxyalkanoates) — a fifth class of physiologically important organic biopolymers. *Angew. Chem. Int. Edn Engl.* **32**, 477–502 (1993).
78. Eerhart, A. J. J. E., Faaij, A. P. C. & Patel, M. K. Replacing fossil based PET with biobased PEF; process analysis, energy and GHG balance. *Energy Environ. Sci.* **5**, 6407–6422 (2012).
- This paper describes a life-cycle assessment that compares the outputs associated with petrochemical-derived PET and biomass-derived PEF.**
79. Burgess, S. K. *et al.* Chain mobility, thermal, and mechanical properties of poly(ethylene furanoate) compared to poly(ethylene terephthalate). *Macromolecules* **47**, 1383–1391 (2014).
80. Jong, E. d., Dam, M. A., Sipos, L. & Gruter, G.-J. M. in *Biobased Monomers, Polymers, and Materials* Vol. 1105 ACS Symposium Series Ch. 1, 1–13 (American Chemical Society, 2012).
81. Delidovich, I. *et al.* Alternative monomers based on lignocellulose and their use for polymer production. *Chem. Rev.* **116**, 1540–1599 (2015).
82. Jeske, R. C., DiCiccio, A. M. & Coates, G. W. Alternating copolymerization of epoxides and cyclic anhydrides: an improved route to aliphatic polyesters. *J. Am. Chem. Soc.* **129**, 11330–11331 (2007).
83. Longo, J. M., DiCiccio, A. M. & Coates, G. W. Poly(propylene succinate): a new polymer stereocomplex. *J. Am. Chem. Soc.* **136**, 15897–15900 (2014).
84. Hong, M. & Chen, E. Y. X. Completely recyclable biopolymers with linear and cyclic topologies via ring-opening polymerization of γ -butyrolactone. *Nature Chem.* **8**, 42–49 (2016).
85. Myers, D., Cyriac, A. & Williams, C. K. Polymer synthesis: to react the impossible ring. *Nature Chem.* **8**, 3–4 (2016).
86. Xiong, M. Y., Schneiderman, D. K., Bates, F. S., Hillmyer, M. A. & Zhang, K. C. Scalable production of mechanically tunable block polymers from sugar. *Proc. Natl Acad. Sci. USA* **111**, 8357–8362 (2014).
87. Juntaro, J. *et al.* Creating hierarchical structures in renewable composites by attaching bacterial cellulose onto sisal fibers. *Adv. Mater.* **20**, 3122–3126 (2008).
88. Braun, B., Dorgan, J. R. & Hollingsworth, L. O. Supra-molecular ecobionanocomposites based on polylactide and cellulosic nanowhiskers: synthesis and properties. *Biomacromolecules* **13**, 2013–2019 (2012).
89. Goffin, A.-L. *et al.* From interfacial ring-opening polymerization to melt processing of cellulose nanowhisker-filled polylactide-based nanocomposites. *Biomacromolecules* **12**, 2456–2465 (2011).
90. Söderqvist Lindblad, M., Albertsson, A. C., Ranucci, E., Laus, M. & Giani, E. Biodegradable polymers from renewable sources: rheological characterization of hemicellulose-based hydrogels. *Biomacromolecules* **6**, 684–690 (2005).
91. Jung, Y. H. *et al.* High-performance green flexible electronics based on biodegradable cellulose nanofibril paper. *Nature Commun.* **6**, 7170 (2015).
92. Haba, O., Tomizuka, H. & Endo, T. Anionic ring-opening polymerization of methyl 4,6-O-benzylidene-2,3-O-carbonyl- α -D-glucopyranoside: a first example of anionic ring-opening polymerization of five-membered cyclic carbonate without elimination of CO₂. *Macromolecules* **38**, 3562–3563 (2005).
93. Mikami, K. *et al.* Polycarbonates derived from glucose via an organocatalytic approach. *J. Am. Chem. Soc.* **135**, 6826–6829 (2013).
94. Upton, B. M. & Kasko, A. M. Strategies for the conversion of lignin to high-value polymeric materials: review and perspective. *Chem. Rev.* **116**, 2275–2306 (2015).
95. Rahimi, A., Ulbrich, A., Coon, J. J. & Stahl, S. S. Formic-acid-induced depolymerization of oxidized lignin to aromatics. *Nature* **515**, 249–252 (2014).
96. Wang, X. & Rinaldi, R. A route for lignin and bio-oil conversion: dehydroxylation of phenols into arenes by catalytic tandem reactions. *Angew. Chem. Int. Edn Engl.* **52**, 11499–11503 (2013).
97. Mialon, L., Pemba, A. G. & Miller, S. A. Biorenewable polyethylene terephthalate mimics derived from lignin and acetic acid. *Green Chem.* **12**, 1704–1706 (2010).
98. Maiorana, A. *et al.* Structure property relationships of biobased *n*-alkyl bisferulate epoxy resins. *Green Chem.* <http://dx.doi.org/10.1039/C6GC01308B> (2016).

Acknowledgements The UK Engineering and Physical Sciences Research Council (EP/K035274/1, EP/M013839/1, EP/L017393/1 and EP/K014070/1) and the China Scholarship Council Imperial Scholarship (Y.Z.) are acknowledged for funding.

Author information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: see go.nature.com/2f1ghz8. Readers are welcome to comment on the online version of this paper at <http://go.nature.com/2f1ghz8>. Correspondence should be addressed to C.K.W. (charlotte.williams@chem.ox.ac.uk).

Reviewer information *Nature* thanks the anonymous reviewers for their contributions to the peer review of this work.

Polymers with autonomous life-cycle control

Jason F. Patrick¹, Maxwell J. Robb^{1,2}, Nancy R. Sottos^{1,3}, Jeffrey S. Moore^{1,2} & Scott R. White^{1,4}

The lifetime of man-made materials is controlled largely by the wear and tear of everyday use, environmental stress and unexpected damage, which ultimately lead to failure and disposal. Smart materials that mimic the ability of living systems to autonomously protect, report, heal and even regenerate in response to damage could increase the lifetime, safety and sustainability of many manufactured items. There are several approaches to achieving these functions using polymer-based materials, but making them work in highly variable, real-world situations is proving challenging.

The life cycle of plastics and other materials used for engineering begins with the extraction of raw materials, followed by the synthesis and processing of the polymer building blocks, which are manufactured into a product that has a particular use or function. The product is eventually degraded or damaged during use, and is ultimately disposed of or recycled^{1,2}. Polymers and polymer-based composites are designed and manufactured to be as robust as possible for a given application, but failure is eventually inevitable. In the case of high-volume and simple, low-cost products, such as the ubiquitous plastic bag, the materials will ideally be recycled after use. But in many instances, the life cycle of polymeric materials could be expanded by programming them with biologically inspired, autonomous functions to protect them from, and to limit, damage, or even to reverse damage and regenerate in response to environmental stress. The attraction of this approach is not only waste reduction, but also the ability to create products with increased safety and superior reliability — a particularly appealing feature for applications such as medical implants, undersea pipelines or structures in space, where damage is difficult to detect, and repair is costly or even impossible.

Research that encompasses chemistry, polymer science, processing and engineering has delivered polymeric materials that have remarkable self-healing, sensing and reporting properties. In this Review we sketch our vision of how such functions could extend the life cycle of functional polymeric materials, and outline the basic performance criteria and material-design principles that should guide the development of practically useful systems. We then examine in more detail the different biologically inspired functions that have been realized, and explore how they can endow materials and devices with improved performance. Finally, we discuss the problems that need to be overcome for this class of polymeric materials to fulfil its promise and find commercial uses.

Polymers with autonomous functions

Nature provides many inspiring examples of materials that perform well in difficult environments, and that can self-heal to regain function when they are damaged. Figure 1 shows the autonomous functions that would enable a polymeric system to similarly maintain and recover its performance throughout its functional life: self-protection guards against potentially damaging environmental factors, such as mechanical stress, chemical corrosion or extreme temperatures; self-reporting capabilities ensure that loss of performance caused by a detrimental event is registered, communicated and ideally initiates action to recover

performance; and when a system has been damaged, self-healing recovers the performance and thereby promotes longevity. However, every polymeric material will inevitably reach an irrecoverable state where self-healing is no longer viable. If such a state is reached because of damage that has physically displaced mass (such as chips, punctures or impacts), it is in principle possible to restore performance by regenerating the lost material. When this is not an option, a desirable final life-cycle step is controlled degradation — achieved by adjusting the rate of degradation or by actively triggering it — to enable active management of the end-of-life removal of the material. For medical implants, this is known as transience³ and ensures that devices function over the required time frame before being resorbed by the body. In other consumer products, controlled degradation can help to recycle material building blocks for use in regeneration or manufacture.

The design of autonomous polymers for life-cycle control is a complex interplay of both intrinsic and extrinsic factors that correlate with the length scale of the damage suffered by a material. The damage length scale (see Fig. 2) is extrinsically influenced by the type of damage event (ballistic impact or fatigue loading), and intrinsically by the inherent nature of the material (the failure and extent of damage of a soft rubber will differ from that of a stiff and brittle polymer). The damage length scale also affects the response function that can be used for repair. For example, dealing with damage that has resulted in the physical removal of mass requires the transport of new material to the damage site for regeneration. This can be achieved by capillary flow on the micrometre scale, but is not practical on the macro scale.

The design of autonomous polymers must also be framed by the property or function that is to be restored. The specifications will be quite different if the target function is mechanical load-carrying capability as opposed to electrical conductivity, for example. Irrespective of function and performance requirements, attempts to put polymers with autonomous functions to practical use also face the considerable challenge of having to do so with a simple, scalable and cost-effective design that meets increasingly stringent regulations.

Figure 2 depicts the three primary approaches to imparting polymers with autonomous function, and gives an indication of the length scales on which they operate. Damage caused by bond scission can be repaired by reversible chemical interactions if the fracture interfaces remain in intimate contact. Success requires rapid bond reorganization at the molecular level, and this can be achieved with supramolecular interactions such as ion pairing. The changes that bring about restoration in

¹Beckman Institute for Advanced Science and Technology, ²Department of Chemistry, ³Department of Materials Science and Engineering, and ⁴Department of Aerospace Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

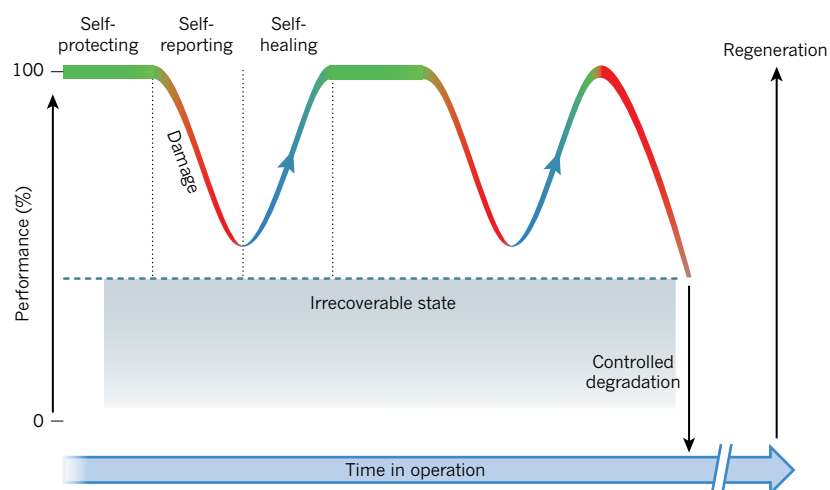


Figure 1 | The life cycle of polymers with autonomous healing functions. Self-protection maintains the structural integrity of a material and minimizes environmental degradation. Self-reporting communicates a loss of performance induced by damage or degradation. Self-healing enables recovery from damage. Controlled degradation brings about the disintegration of a material after irrecoverable damage has occurred or after a predetermined time in operation. Regeneration effectively restarts the polymer life cycle by rebuilding the material.

this case take place internally and require molecular engineering of the polymer to equip it with the intrinsic characteristics of reconfigurable or dynamic bonds. When the damage grows to the microscale, healing by dynamic bonding is usually no longer viable, and an extrinsic source of repair agents is needed. In such instances, two basic strategies are available: the incorporation in the polymeric material of mesoscale additives (such as microcapsules or fibres) with a functional fluid payload that is discharged and initiates healing once damage occurs; or the incorporation of a larger-scale, vascular fluid-delivery network. Healing is initiated either when repair agents are mixed, or when a repair agent interacts with an additional component present in the matrix (such as a catalyst). Both of these trigger chemical reactions that transform the fluid into a structural adhesive that bonds the microcrack interfaces together. In cases where damage has resulted in the physical displacement of mass (for example, puncture), and a substantial volume of repair agents is needed, healing is only feasible using delivery from a vascular network.

The three approaches to autonomous polymers sketched in Fig. 2 address the spectrum of damage from ångström-scale bond scission to microcracking and millimetre-scale puncture damage. They have enabled the development of systems that exhibit self-protection, self-reporting, self-healing, regeneration and controlled degradation (see Fig. 1), mainly in proof-of-principle demonstrations, although some are nearing commercialization.

Self-protection

The best way of ensuring the longevity of materials and systems is to prevent damage before it occurs, and coatings have long been used for that purpose. Coatings passively protect substrates against mechanical and chemical degradation by forming a physical barrier that increases resistance to corrosion or wear. Augmenting this passive protection with self-repairing capabilities improves the overall performance of coatings⁴ (Fig. 3a). This additional protection is typically achieved by incorporating mesoscale additives that deliver active payloads consisting of corrosion inhibitors⁵ or compounds that react with water to form a physical⁶ or hydrophobic⁷ protective barrier (Fig. 3b) in response to surface cracking, ablation or corrosion^{8–10}. Purely inorganic sol–gel coatings, for example, have been shown to achieve enhanced corrosion protection of an aluminium alloy¹¹ by incorporating mesoporous silica nanoparticles loaded with a corrosion inhibitor. Protection strategies that aim to create repellent surface layers rely on tailored surface chemistry. Although it is a departure from our theme of polymers with autonomous functions, another approach to creating self-protecting functional surfaces relies on liquid-repellent surfaces comprising porous substrates infused with a lubricating liquid. The porous substrate locks the lubricant in place, giving a smooth, defect-free and stable liquid surface that can be tailored to repel a wide range of different liquids and that readily self-repairs after mechanical damage^{12,13}.

As well as mechanical damage and corrosion, materials and devices also frequently face the threat of thermal degradation. Consider, for example, lightweight, fibre-reinforced, thermoset polymer composites, which are inherently resistant to many forms of environmental deterioration (such as corrosion). They are increasingly replacing traditional structural materials in advanced engineering applications, but their mechanical performance rapidly diminishes at temperatures near the glass transition temperature of the polymer matrix (typically at or below 200 °C). At temperatures above the glass transition, the matrix can no longer carry structural loads. The development of matrix materials with enhanced thermal stability is one potential solution for expanding the scope of polymer fibre composites.

Alternatively, circulating fluids through vascular networks embedded within the composite, which occupy a small fraction of the overall material volume^{14,15}, can protect against thermal degradation while retaining structural performance^{16–18} — an approach well known in the aerospace industry and long used to cool high-performance ceramic turbine blades¹⁹. Indeed, a recent study²⁰ showed that an actively cooled vascular polymer matrix composite, with a matrix glass transition temperature of 150 °C and a channel volume comprising only a few per cent, retains its flexural stiffness upon continuous exposure to 325 °C. When endowed with such active cooling functionality, it should in principle be possible to use conventional thermoset, polymer matrix composites in applications subject to high thermomechanical loading, such as aerospace structures, microelectronics packaging, or battery packaging for electric vehicles. Putting self-cooling materials to practical use, however, will require efficient and inexpensive means of manufacture, and innovative approaches for accommodating the coolant and pumping equipment they require.

Self-protection implemented at the device level is exemplified by high-energy-density lithium-ion batteries that manage thermal runaway, which poses a significant safety hazard after batteries have been damaged or discharged too quickly. One strategy is to incorporate thermally responsive polymer microspheres that melt above a critical temperature and disrupt conduction pathways within the battery to irreversibly shut down the cell²¹. Reversible shutdown, in which normal battery operation is restored once the thermal perturbation is removed, can be achieved using conductive particles within a temperature-responsive polymer binder²². Thermal degradation, however, is not the only operational challenge for high-energy-density batteries. Mechanical damage resulting from dramatic volumetric expansion and contraction in next-generation silicon electrodes leads to rapid capacity fade and diminished cyclability. This problem can be mitigated by using silicon particles in combination with a flexible polymer binder that accommodates large volume change through reversible hydrogen bonding, significantly improving the cell's lifetime and retention of capacity²³.

The examples so far illustrate protection of functional performance at a system or device level, but molecular-scale protection can also

be achieved by using mechanochemically active polymers²⁴. These materials incorporate force-sensitive molecules, known as mechanophores, that directly harness mechanical energy to promote productive chemical transformations²⁵. Mechanical stress usually breaks covalent bonds, and thereby degrades a material's properties and performance, but mechanophores give materials the ability to survive otherwise degrading conditions. This ability could be useful when operating conditions are unpredictable and the use of alternative materials with overengineered properties would be wasteful and inefficient. A recent example of mechanochemical self-protection used polymers containing dibromocyclopropane mechanophores, which transform under shear stress into reactive functional groups that can be crosslinked *in situ* to bring about autonomic strengthening²⁶ and enhance stiffness. Mechanophores with a wide range of chemical functionalities are becoming available^{27,28}. Noteworthy in this context is the mechanical activation of catalysts²⁹, which makes it possible to initiate a variety of protective chemistries (such as polymerization or crosslinking reactions) while also achieving chemical amplification. However, these materials are difficult to synthesize and are expensive, which could limit their practical application, but there may be opportunities, particularly when coupling mechanochemically active polymers with mesoscale additives to generate materials with hierarchical function. For example, polymers that produce acid in response to mechanical stress³⁰ could be combined with a pH-responsive delivery vehicle to give a system that will release its payload under mechanical force.

Self-reporting

Many systems and devices benefit from the ability to autonomously signal the occurrence of stress or damage (Fig. 3c). A powerful strategy for creating such a warning system uses mechanophores to monitor molecular force and convey visual information about the condition and mechanical history of a polymer. As with the mechanophores that bring about molecular self-protection, the mechanical forces that lead to damage induce a chemical transformation in the mechanophore, but

to enable self-reporting, this chemical transformation should be accompanied by changes in optical properties. For example, the mechanically activated ring-opening reaction of spiropyran³¹ in bulk polymeric materials results in changes in both the visible colour (from yellow to red) and fluorescence. In many cases, mechanochemical transformations are achieved at the expense of permanent plastic deformation of the material, which exemplifies the difficulty of achieving efficient transduction of mechanical energy to a specific mechanophore molecule. Integrating a spiropyran mechanophore into a silicone elastomer³², in contrast, provides a visually discernible indication of stress or strain in combination with complete shape recovery (Fig. 3d). Three-dimensional (3D) printing of mechanochromic polymers is also possible, and has been used to produce a prototype force sensor that can be used to evaluate loads simply by observation of the mechanically induced colour change³³. These examples all require a light source for visualization, but mechanophores such as bis(adamantyl)-1,2-dioxetane can also use mechanical energy to generate light that is visualized directly³⁴. An earlier approach to molecular strain sensors used dye aggregates in polymer blends, in which macroscopic deformation affects the aggregation and thereby produces changes in the photoluminescence emission colour³⁵.

Self-reporting of mechanical damage is also possible using mesoscale additives and vascular networks³⁶ that, when ruptured, release indicators for visual detection through a change in colour or fluorescence³⁷. Most systems rely on a chemical reaction between an indicator molecule and a secondary reagent or catalyst incorporated into the polymeric material, or through the use of dual-capsule systems containing the two reactive precursors³⁸. For example, a visual colour change is generated in regions of mechanical damage using microcapsules containing a colourless conjugated cyclic monomer that when released and allowed to react with an embedded catalyst transforms into a deeply coloured polymer³⁹. Microcapsules containing a pH-sensitive dye⁴⁰ that reacts with residual amine groups in an epoxy polymer matrix have led to particularly robust and clear visual indications of damage even on the micrometre scale, but this approach shares the limitation of other systems of requiring

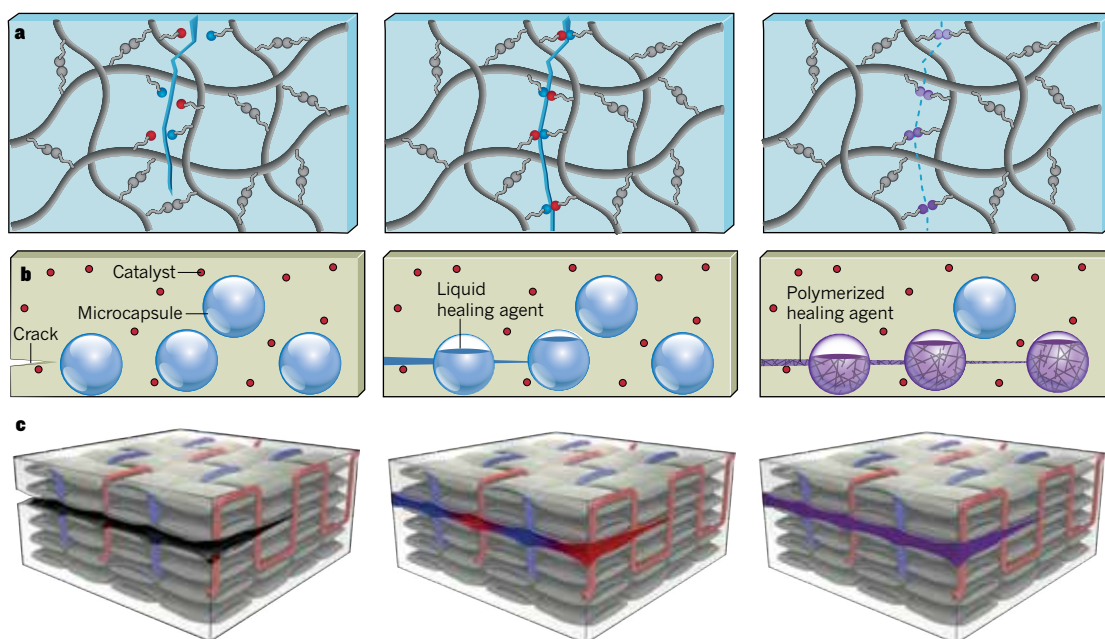


Figure 2 | Multiscale strategies for autonomous repair functions in polymeric materials. Three primary strategies enable autonomous repair functions in polymeric materials. They are exemplified here by self-healing over a range of damage length scales. **a**, At the smallest scale (ångströms), molecular engineering brings about the repair of damage by dynamic or reversible bonding of fractured interfaces that are in intimate contact. Bond scission occurs (left) followed by dynamic rebonding (middle), and the fracture is mended (right). **b**, Mesoscale additives that are dispersed in a

polymeric material can store and release healing agents to recover damage at the microscale (<100 μm). A microcrack occurs (left), the microcapsules release a healing agent (middle) that polymerizes on contact with an embedded catalyst, and the crack is healed (right). **c**, Vascular networks capable of fluid circulation and repeated delivery of healing agents can repair material from the micrometre to the millimetre scale. Delamination occurs (left), the ruptured vascular network releases reactive, liquid healing agents (middle), and the delamination is repaired (right).

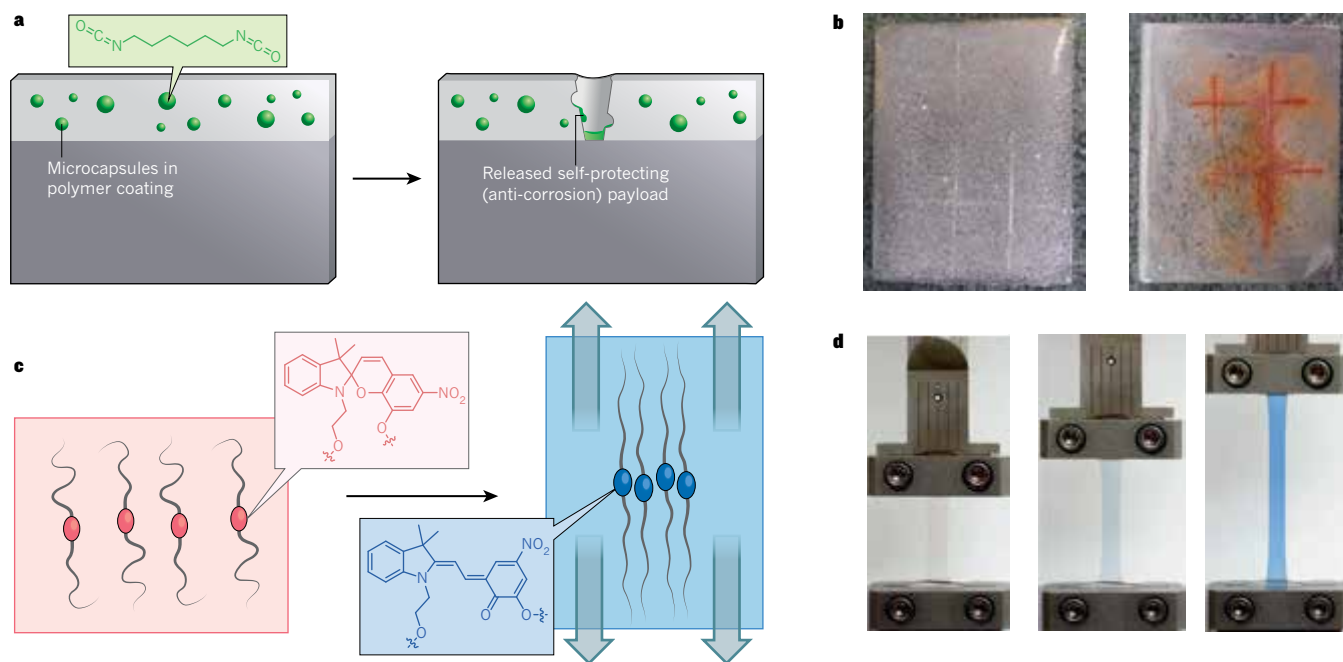


Figure 3 | Autonomous prevention and communication of material degradation. **a**, A self-protecting material incorporating mesoscale additives. Damage to a coating causes microcapsules to rupture and release an anti-corrosion payload that protects the underlying substrate. **b**, Examples of coated steel substrates subjected to corrosive environments. Left, an autonomous self-protecting epoxy coating incorporating microcapsules with a reactive isocyanate payload. Scratch damage to the coating releases the capsule payload and protects the underlying steel substrate from active corrosion. Right, a

conventional epoxy coating (with no microcapsules) in which scratch damage initiates widespread corrosion of the substrate (red). Images reprinted with permission from ref. 6. **c**, A self-reporting material incorporating a molecular force probe (mechanophore). Mechanical activation causes the material to change colour, providing a visual indication of critical stress or strain. **d**, A polydimethylsiloxane elastomer containing a spiropyran mechanophore that turns blue during tensile loading, providing a visual indication of mechanical stress or strain. Images reprinted with permission from ref. 32.

specific chemical interactions. This highlights the need for more general visualization strategies that are applicable to a wider range of materials, irrespective of chemical composition. With this objective in mind, a recently developed approach uses microcapsules containing molecules that exhibit aggregation-induced emission, which become fluorescent after rupture and release by a physical change of state⁴¹. Another example of a general, autonomous approach for damage detection is the entropy-driven migration of fluorescent nanoparticles to cracks in layered composite structures⁴². Compared with mechanophore activation, which typically requires large and often irreversible polymer deformation, microcapsules produce a more permanent response with enhanced sensitivity to microscale damage.

Self-healing

An ideal self-healing polymer system would repair any damage it suffers in a site-specific and fully autonomous fashion to regain functional performance. Of the three approaches outlined in Fig. 2, self-healing based on molecular engineering of the polymer comes closest to this ideal because repair can occur wherever it is needed without the use of additives^{43,44}. However, many early examples of this sort required external energy input to drive the reorganization of bonds and polymer chains necessary for the repair process. One such system comprising a covalently crosslinked thermoset polymer based on Diels–Alder chemistry is shown in Fig. 4a,b. The thermally reversible nature of the covalent bonds in this system enables dynamic exchange when heated to mend cracks⁴⁵. Such thermal re-mending has also been accomplished with thermoplastic phases that are incorporated into a crosslinked polymer matrix to enable repair of cracks through local entanglement when heated⁴⁶. Vitrimers⁴⁷ are a similar class of polymer that become malleable and mendable at high temperatures but retain their covalently crosslinked network throughout⁴⁸, so the system's viscosity changes gradually, allowing local mending and reshaping. Light has also been used to induce the mobility⁴⁹ needed for healing, and to allow localized heating and

self-healing in metallo-supramolecular polymers⁵⁰. When embedding super-paramagnetic nanoparticles, even an oscillating magnetic field can induce amorphous flow and re-mending in a thermoplastic polymer⁵¹ by means of the rapidly vibrating particles.

But for self-healing to be fully autonomous, it needs to proceed without the input of external energy. The molecular-engineering approach using dynamic bonding accomplishes this relatively readily in soft, rubbery polymers or gels. A striking example, based on hydrogen-bonding interactions between small molecules, is a highly extensible rubber that is capable of autonomously, repeatedly and fully healing fractured surfaces at room temperature⁵². Similar design principles have also produced self-healing thermoplastic elastomers⁵³ with a mechanical performance approaching that of conventional, structural polymeric materials. The interaction of charged ionic species is an alternative bonding motif that can yield self-repairing rubbers⁵⁴ and hydrogels. When implemented using polymer-modified clay nanosheets and a dendritic binder⁵⁵, the resulting hydrogels retain their shape and completely recover their mechanical integrity after cleavage and reassembly. Although such supramolecular interactions are an obvious bonding platform, dynamic covalent chemistry has also been used in adaptive⁵⁶ and self-healing polymers^{57,58}. The ease with which self-healing proceeds, and the ability to heal repeatedly, are attractive features of the molecular-engineering approach, but it is largely restricted to elastomeric polymers and can usually deal with only small damage length scales, as there needs to be intimate contact between damaged surfaces.

As early, inspirational work showed⁵⁹, some limitations on materials and damage length scales can be overcome by using extrinsic healing agents embedded in encapsulated form in the material, ready to be released when damage occurs. An important proof-of-principle demonstration of such autonomous self-healing in a structural polymer introduced the basic and now widely used microcapsule design⁶⁰. Crack propagation ruptured the microcapsules contained within the material and released the encapsulated monomer fluid. When this

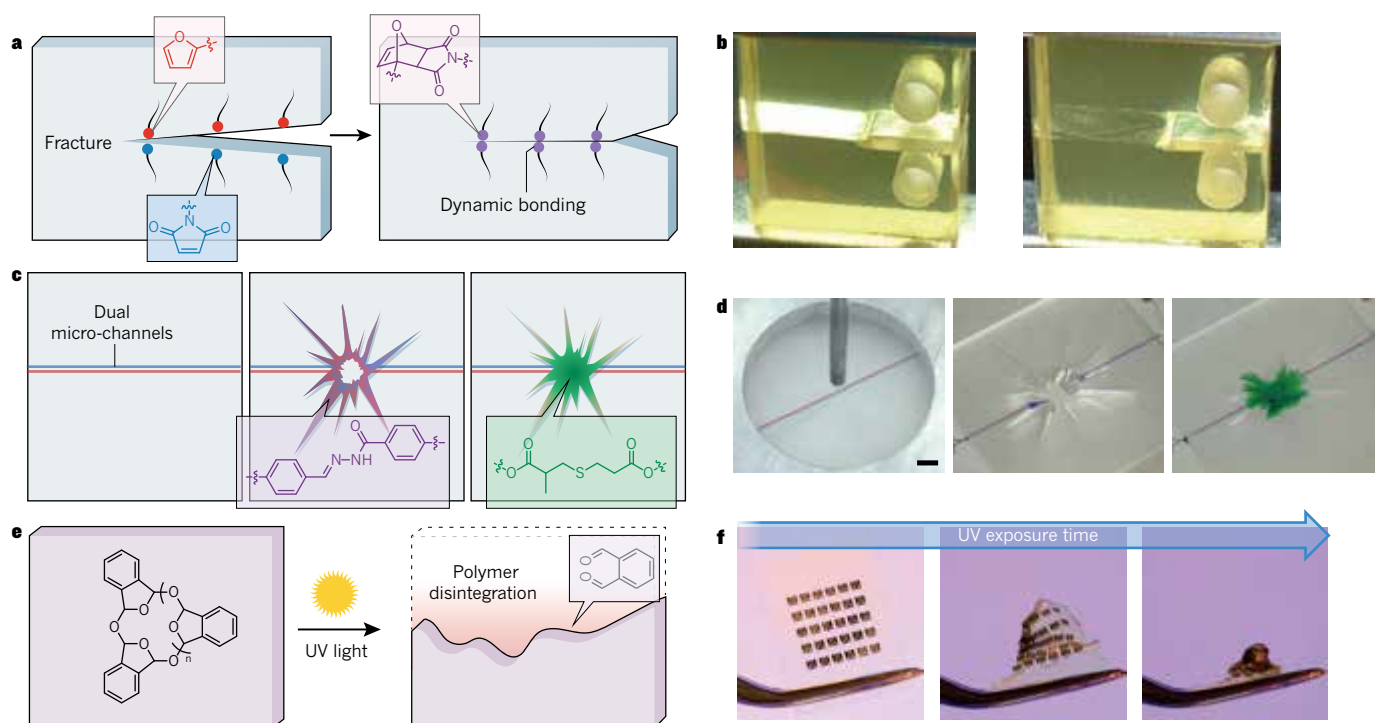


Figure 4 | Autonomous healing, regeneration and degradation. **a**, A self-healing material using a molecular-engineering strategy. Reversible bonding between the two crack surfaces enables repair after they are brought into intimate contact. **b**, Repair of fracture damage in a re-mendable, covalently crosslinked polymeric material after heating. Left, the fracture plane can be clearly seen because of light reflection. Right, after clamping the fracture and heating to promote repair, the crack is no longer visible. Images reprinted with permission from ref. 45. **c**, Regeneration of a polymeric material. Micro-channels embedded in the polymer (left) deliver liquid healing agents

to the site of large-scale damage (middle) to restore lost material (right). **d**, Dual microvascular networks (blue and red; left) deliver a two-part healing system to restore puncture damage in an epoxy substrate⁸⁶. The puncture (~10 mm; middle) is completely filled with the healing agents and closed after pressurized delivery (right). **e**, A material programmed for controlled degradation (left). An environmental stimulus (such as sunlight) triggers the degradation of the material (right). **f**, Packaging of an electronic device with a transient polymer enables the controlled disintegration of the device (left to right) when exposed to UV light. Reprinted with permission from ref. 89.

fluid made contact with catalyst particles contained in the crosslinked epoxy matrix, it polymerized to bond the fractured surfaces and restore mechanical performance. This example motivated the development of many capsule-based self-healing systems, implemented in both hard and soft polymers and using a multitude of different healing chemistries (for in-depth reviews, see refs 43 and 61). Importantly, microcapsules can be incorporated into polymers without sacrificing the inherent fracture toughness and related mechanical properties^{62,63}.

Critical to the further success and translation of these systems is the need for robust capsules and healing chemistries that remain stable and functional, despite being exposed to a variety of mechanical, thermal and chemical environments, until damage occurs. Protective capsule coatings can enhance the stability of microcapsules that are exposed to challenging processing and environmental conditions⁶⁴. However, this still leaves one major limitation: capsules can be used only once in a particular location to release their liquid payload and initiate healing. One possible way of overcoming this issue is to use capsules with semi-permeable shells⁶⁵ for tunable and controlled payload release.

Although microcapsules can be made to be stable and functional, and incorporated using standard material-processing techniques, a fundamental limitation is the total amount of healing agent that can be delivered to damaged areas. Vascular networks overcome these shortcomings and also allow repeated healing by replenishing the supply of healing agent to sites of repeated damage. Precursors of this basic strategy (the use of one-dimensional tubular structures for delivering healing agent) involved isolated hollow, glass fibres³⁶ or microchannels^{66,67}. Multiple cycles of self-healing have been achieved in both coatings⁶⁸ and neat polymers⁶⁹ by using interconnected, three-dimensional microvascular networks. More-complex, interpenetrating 3D microvascular networks containing two-part healing agents have been shown to sustain more

than 30 cycles of repeated healing⁷⁰, and adding a third independent network enabled regulation of the healing temperature and time⁷¹, which could, in principle, enable healing under otherwise prohibitive environmental conditions.

The development of vascular-based healing has relied on the efficient integration of complex networks into high-performance, fibre-reinforced composites. Early fabrication techniques^{72,73} either produced only one-dimensional isolated channels with limited fluid pathways, or used direct-write processes⁷⁴ that are incompatible with most industrial manufacturing techniques. The more recent approach of using a robust, sacrificial template of polylactic acid polymer can be seamlessly integrated into existing manufacturing processes for polymer-based composites^{14,15}. This so-called vaporization of sacrificial components technique has enabled more complex vascular architectures to be seamlessly integrated into fibre-reinforced composites with no decrease in mechanical strength and stiffness¹⁷, while increasing delamination resistance and enabling multiple cycles of healing at high efficiencies⁷⁵. Many significant hurdles remain to be overcome before such self-healing, high-performance composites can be put to practical use, however. Not only are costs potentially prohibitive, but the delivery of healing agents needs to become fully autonomous, network design must be made redundant to damage and/or blockage, and *in situ* repair of ruptured vasculature must be realized to achieve sustained delivery of agents and prolonged healing cycles.

The biological world has inspired many of the concepts that guide the development of materials and systems capable of autonomously repairing damage, but self-healing remains relatively unexplored by the biomaterials research community. Toxicity and biocompatibility impose important constraints that have so far restricted progress to proof-of-concept studies^{76–78} demonstrating the feasibility of

microcapsule-based self-healing of bone cement (a polymeric material used to anchor implants including artificial joints, which can lose function through wear and microcracking). Further development of these and other synthetic polymer-based biomaterials with the ability to autonomously repair damage holds considerable promise for future biomedical applications⁷⁹.

The examples of self-healing discussed above all relate to the restoration of mechanical performance, but the approach based on mesoscale additives in particular is well suited for targeting other functions as well. Efforts so far have focused on electrical properties (either on their own^{80–83} or in conjunction with the restoration of mechanical properties⁸⁴ that might be useful for electronic skin applications), but we envisage that materials with the ability to repair other functional properties will also be developed.

Regeneration

The ultimate challenge in the development of self-healing materials is regeneration — the ability to replace severely damaged or lost materials (analogous to the biological regeneration of tissues). Damage that involves a loss of mass on this scale (Fig. 4c) requires a delivery system to supply sufficient amounts of healing agents. As with biological growth and regeneration⁸⁵, this requires a rich interplay of transport, kinetics and dynamically changing material properties. This concept is illustrated by the restoration and healing of millimetre-scale puncture damage in an epoxy polymer⁸⁶, which used the vascular delivery of two-stage healing chemistry to combat the constraints of gravity and surface tension on the amount of healing fluid that can be retained in a large void. The first healing stage involved the gelation of a monomer to form a permeable scaffold that allowed the accumulation of material to completely fill the damaged region. The second healing stage involved the polymerization of a secondary orthogonal monomer to generate a robust structural polymer (Fig. 4d). This strategy proved effective at filling the central puncture, but it was unable to ensure complete infiltration and repair of the smaller radial cracks that resulted from the impact. This shortcoming aptly illustrates the problems posed by repairing damage that spans multiple size scales simultaneously.

Controlled degradation

Even the most robust self-healing materials will eventually reach the end of their life. Ideally, there will be controlled degradation of the material to recover useful products for recycling (Fig. 4e). Controlled degradation is not only desirable when damage has led to an irrecoverable state in which self-healing and regeneration are no longer viable, but also when a material or device is needed for only a limited time period, so its transience, rather than its stability, is desirable. Of course, degradability (or transience) has been pursued for some time, not least to deal with the millions of tonnes of plastic waste that is discarded every year. But although materials such as biodegradable bags and other personal products are becoming widespread, attempts to make advanced polymer-based materials and devices degradable are still at an early stage.

One approach to the paradoxical materials-design challenge of obtaining readily degradable materials that are robust and stable while they are being used is to design materials that will disintegrate when exposed to a particular triggering event. One promising example uses poly(benzyl ether) as a basic platform that can easily be modified to produce traditional polymeric materials, leading to chemically induced depolymerization for selective end-of-life recycling^{87,88}. In the case of derivatives of poly(phthalaldehyde), which are used as packaging material for electronic devices, degradation has been triggered by ultraviolet light⁸⁹ and heat⁹⁰. But there remains a need for broader classes of materials that can readily degrade upon exposure to suitable stimuli.

Transience also arises naturally in cases where materials continuously disintegrate during use. This feature can be controlled to ensure that the degradation rate is such that adequate functional performance is maintained over the intended period of use. The resorbable sutures often used by surgeons illustrate the benefits and elegance of this approach.

A more recent example⁹¹ (involving both organic and inorganic materials) is the use of multifunctional silicon nanomembrane sensors for the brain that are implanted in rats, as these can be resorbed naturally to avoid surgical removal (Fig. 4f). Especially in the case of medical implants that have active functions, accelerated degradation when exposed to a suitable stimulus — so-called triggered transience — could further expand the opportunities for use.

Although some strategies make it possible to achieve controlled degradation, a more demanding objective is the recapture and reuse of the resulting products. Proof-of-principle experiments have demonstrated the mechanically triggered depolymerization of poly(phthalaldehyde) into monomer that was subsequently repolymerized⁹². But major challenges remain before a fully integrated materials system can be created that approaches the ideal of a closed, autonomous, polymer life cycle.

Outlook

Materials have traditionally been developed to be robust enough to withstand the wear and tear of normal use. This approach has worked well, but moving towards active damage and life-cycle management using autonomous functions promises materials with enhanced, safer and more efficient performance, while minimizing resource use and waste production. There are several major obstacles to using autonomous polymers in real-world applications, not least the fact that they have so far largely been developed and tested in highly controlled and optimized laboratory settings. In stark contrast, materials in service will need to perform in highly variable environmental conditions and remain stable throughout their operational life. The chemistry that enables healing in particular is often sensitive to temperature, humidity, pressure, pH and atmospheric oxygen, which all make it difficult to maintain robust healing performance. If autonomous technologies are to enter the commercial sector, it is essential that stability is established over periods much longer than the six months explored in one study⁹³. We also note that many healing agents are not only expensive but also toxic, which largely precludes their commercial use. In addition, the need for self-reporting and self-healing to occur quickly enough to deal with damage — particularly if it rapidly worsens or propagates — poses significant challenges. Most of the systems we have discussed require lengthy healing periods that often cannot be readily accommodated, especially if healing under continuous loading is not effective.

The commercialization of self-healing polymer-based materials faces significant practical challenges, and developments in the field so far have been largely empirically driven owing to the relative lack of suitable computational tools and models. We envisage that efforts to redress this situation will deliver considerable benefits. For example, computer modelling can provide deeper insight into the current behaviour of autonomous polymers, and thereby aid the formulation of effective guidelines for optimizing both the synthesis of these polymers and the design of the system. Multi-scale models are needed that effectively connect chemical transformations on the molecular scale with relatively local material responses on the mesoscale, and with the macroscale behaviour of the system overall. Modelling will also need to deal with a wide range of processes and phases, given that structural and other damage gives rise to a healing response in which the flow of repair agents and chemical reactions involving fluid and solid phases has a critical role. Steps towards modelling materials and systems with autonomous functions have been taken^{94–99}, but it is difficult to capture sufficient detail about the underlying mechanisms to enable meaningful predictions, while still ensuring that model calculations can be performed on reasonable timescales.

Even when an effectively functioning, self-healing material or system is available, practical implementation in the commercial sector will require the product to be manufactured in commercially relevant quantities and deliver value in the marketplace. This remains a largely unsolved problem for autonomous polymers. Some soft, rubbery polymers with fully autonomous self-repairing properties can be produced using simple processes and inexpensive, renewable materials⁵². When

aiming for structural applications, however, materials usually require extrinsic healing agents. Systems that rely on microcapsules for this purpose face the problem that although emulsion processes can produce reasonable quantities of capsule material, they are restrictive in terms of the chemistry that can be used. For example, hydrophilic healing agents cannot be encapsulated in an oil-in-water emulsion process (which is by far the most common commercial encapsulation method), and reactive core materials are chemically incompatible with many polymer shell-forming chemistries. Microfluid-based processes¹⁰⁰ can overcome some of these problems but are unlikely to be cost-effective on a commercial scale owing to difficulties of scalability and the slow rate of capsule production. Another obstacle for systems that use mesoscale additives, especially those carrying sensitive payloads, is their incompatibility with the processing methods used to manufacture high-performance fibre-composite materials. Autonomous systems that use vascular networks have largely overcome these problems, but the complexity of their design, and the need to introduce and pump reactive agents within the network, pose further manufacturing and implementation challenges. Advances in additive manufacturing for the rapid production of 3D scaffolds¹⁰¹ or even 4D-printing concepts¹⁰² may overcome some of these hurdles, but commercial viability is unlikely in the near future except in specialized applications in which the additional costs are warranted.

The ultimate goal for autonomous polymers is to perform the desired function for as long as necessary without being replaced or requiring external maintenance. The end of life should be designed in such a way that the building blocks of the polymer can be recaptured and recycled efficiently, reducing the amount that is discarded in landfill. Imagine a car tyre that lasts the entire lifetime of the vehicle, before being recycled and regenerated with 100% efficiency. Alternatively, a bridge that is painted once and is protected from corrosion for its entire lifetime will greatly reduce maintenance costs while improving the safety of our civil infrastructure. Beyond simply replacing existing materials, self-healing polymers might also enable product designers to explore new concepts that incorporate healing as an integral part of the design. Irrespective of whether and how the futuristic goal of autonomous control of the entire polymer life cycle can be achieved, the first exciting steps have been taken, and the challenge for the field now is to deliver on the promise of improved sustainability by providing smarter, safer, better-performing and longer-lasting materials.

Received 11 March; accepted 12 October 2016.

- Rebitzky, G. *et al.* Life cycle assessment: part 1: framework, goal and scope definition, inventory analysis, and applications. *Environ. Int.* **30**, 701–720 (2004).
- Hopewell, J., Dvorak, R. & Kosior, E. Plastics recycling: challenges and opportunities. *Phil. Trans. R. Soc. B* **364**, 2115–2126 (2009).
- Hwang, S.-W. *et al.* A physically transient form of silicon electronics. *Science* **337**, 1640–1644 (2012).
This paper describes the first physically transient electronics and application in an implantable biomedical device.
- García, S. J., Fischer, H. R. & van der Zwaag, S. A critical appraisal of the potential of self healing polymeric coatings. *Prog. Org. Coatings* **72**, 211–221 (2011).
- Selvakumar, N., Jeyasubramanian, K. & Sharmila, R. Smart coating for corrosion protection by adopting nano particles. *Prog. Org. Coatings* **74**, 461–469 (2012).
- Huang, M. & Yang, J. Facile microencapsulation of HDI for self-healing anticorrosion coatings. *J. Mater. Chem.* **21**, 11123–11130 (2011).
- Latnikova, A., Grigoriev, D. O., Hartmann, J., Möhwald, H. & Shchukin, D. G. Polyfunctional active coatings with damage-triggered water-repelling effect. *Soft Matter* **7**, 369–372 (2011).
- Shchukin, D. & Möhwald, H. A coat of many functions. *Science* **341**, 1458–1459 (2013).
- Shchukin, D. G. Container-based multifunctional self-healing polymer coatings. *Polym. Chem.* **4**, 4871–4877 (2013).
- Esser-Kahn, A. P., Sottos, N. R., White, S. R. & Moore, J. S. Programmable microcapsules from self-immolative polymers. *J. Am. Chem. Soc.* **132**, 10266–10268 (2010).
- Borisova, D., Möhwald, H. & Shchukin, D. G. Mesoporous silica nanoparticles for active corrosion protection. *ACS Nano* **5**, 1939–1946 (2011).
- Wong, T.-S. *et al.* Bioinspired self-repairing slippery surfaces with pressure-stable omniphobicity. *Nature* **477**, 443–447 (2011).
- Cui, J., Daniel, D., Grinthal, A., Lin, K. & Aizenberg, J. Dynamic polymer systems with self-regulated secretion for the control of surface properties and material healing. *Nature Mater.* **14**, 790–795 (2015).
- Esser-Kahn, A. P. *et al.* Three-dimensional microvascular fiber-reinforced composites. *Adv. Mater.* **23**, 3654–3658 (2011).
- Gergely, R. C. R. *et al.* Multidimensional vascularized polymers using degradable sacrificial templates. *Adv. Funct. Mater.* **25**, 1043–1052 (2015).
- Kousourakis, A., Mouritz, A. P. & Bannister, M. K. Interlaminar properties of polymer laminates containing internal sensor cavities. *Compos. Struct.* **75**, 610–618 (2006).
- Coppola, A. M., Thakre, P. R., Sottos, N. R. & White, S. R. Tensile properties and damage evolution in vascular 3D woven glass/epoxy composites. *Composites A* **59**, 9–17 (2014).
- Hartl, D. J., Frank, G. J. & Baur, J. W. Effects of microchannels on the mechanical performance of multifunctional composite laminates with unidirectional laminae. *Compos. Struct.* **143**, 242–254 (2016).
- Han, J.-C., Dutta, S. & Ekkad, S. *Gas Turbine Heat Transfer and Cooling Technology*, 2nd edn (CRC Press, 2012).
- Coppola, A. M., Griffin, A. S., Sottos, N. R. & White, S. R. Retention of mechanical performance of polymer matrix composites above the glass transition temperature by vascular cooling. *Composites A* **78**, 412–423 (2015).
- Baginska, M. *et al.* Autonomic shutdown of lithium-ion batteries using thermoresponsive microspheres. *Adv. Energy Mater.* **2**, 583–590 (2012).
- Chen, Z. *et al.* Fast and reversible thermoresponsive polymer switching materials for safer batteries. *Nature Energy* **1**, 15009 (2016).
- Wang, C. *et al.* Self-healing chemistry enables the stable operation of silicon microparticle anodes for high-energy lithium-ion batteries. *Nature Chem.* **5**, 1042–1048 (2013).
- Berkowski, K. L., Potisek, S. L., Hickenboth, C. R. & Moore, J. S. Ultrasound-induced site-specific cleavage of azo-functionalized poly(ethylene glycol). *Macromolecules* **38**, 8975–8978 (2005).
- Hickenboth, C. R. *et al.* Biasing reaction pathways with mechanical force. *Nature* **446**, 423–427 (2007).
- Ramirez, A. L. B. *et al.* Mechanochemical strengthening of a synthetic polymer in response to typically destructive shear forces. *Nature Chem.* **5**, 757–761 (2013).
This paper describes in situ strengthening of polymers in response to shear forces by mechanically induced covalent crosslinking.
- Li, J., Nagamani, C. & Moore, J. S. Polymer mechanochemistry: from destructive to productive. *Acc. Chem. Res.* **48**, 2181–2190 (2015).
- Simon, Y. C. & Craig, S. L. (eds). *Mechanochemistry in Materials* (Royal Society of Chemistry, in the press).
- Piermattei, A., Karthikeyan, S. & Sijbesma, R. P. Activating catalysts with mechanical force. *Nature Chem.* **1**, 133–137 (2009).
- Diesendruck, C. E. *et al.* Proton-coupled mechanochemical transduction: A mechanogenerated acid. *J. Am. Chem. Soc.* **134**, 12446–12449 (2012).
- Davis, D. A. *et al.* Force-induced activation of covalent bonds in mechanoresponsive polymeric materials. *Nature* **459**, 68–72 (2009).
This paper describes the first demonstration of a mechanically activated covalent reaction in bulk polymeric materials.
- Gossweiler, G. R. *et al.* Mechanochemical activation of covalent bonds in polymers with full and repeatable macroscopic shape recovery. *ACS Macro Lett.* **3**, 216–219 (2014).
- Peterson, G. I., Larsen, M. B., Ganter, M. A., Storti, D. W. & Boydston, A. J. 3D-printed mechanochromic materials. *ACS Appl. Mater. Interf.* **7**, 577–583 (2015).
- Chen, Y. *et al.* Mechanically induced chemiluminescence from polymers incorporating a 1,2-dioxetane unit in the main chain. *Nature Chem.* **4**, 559–562 (2012).
- Löwe, C. & Weder, C. Oligo(p-phenylene vinylene) excimers as molecular probes: Deformation-induced color changes in photoluminescent polymer blends. *Adv. Mater.* **14**, 1625–1629 (2002).
- Pang, J. W. C. & Bond, I. P. A hollow fibre reinforced polymer composite encompassing self-healing and enhanced damage visibility. *Compos. Sci. Technol.* **65**, 1791–1799 (2005).
This paper describes damage indication and healing of a fibre-reinforced polymer composite by the incorporation of hollow glass fibres that rupture and release reactive liquids.
- van den Dungen, E. T. A., Loos, B. & Klumperman, B. Use of a profluorophore for visualization of the rupture of capsules in self-healing coatings. *Macromol. Rapid Commun.* **31**, 625–628 (2010).
- Lavrenova, A., Farkas, J., Weder, C. & Simon, Y. C. Visualization of polymer deformation using microcapsules filled with charge-transfer complex precursors. *ACS Appl. Mater. Interf.* **7**, 21828–21834 (2015).
- Odum, S. A. *et al.* Visual indication of mechanical damage using core-shell microcapsules. *ACS Appl. Mater. Interf.* **3**, 4547–4551 (2011).
- Li, W. *et al.* Autonomous indication of mechanical damage in polymeric coatings. *Adv. Mater.* **28**, 2189–2194 (2016).
- Robb, M. J. *et al.* A robust damage-reporting strategy for polymeric materials enabled by aggregation-induced emission. *ACS Central Sci.* **2**, 598–603 (2016).
- Gupta, S., Zhang, Q., Emrick, T., Balazs, A. C. & Russell, T. P. Entropy-driven segregation of nanoparticles to cracks in multilayered composite polymer structures. *Nature Mater.* **5**, 229–233 (2006).
- Blaiszik, B. J. *et al.* Self-healing polymers and composites. *Annu. Rev. Mater. Res.* **40**, 179–211 (2010).
- Murphy, E. B. & Wudl, F. The world of smart healable materials. *Prog. Polym. Sci.* **35**, 223–251 (2010).

45. Chen, X. *et al.* A thermally re-mendable cross-linked polymeric material. *Science* **295**, 1698–1702 (2002).
This paper describes multiple cycles of healing in a crosslinked polymer network by thermally reversible covalent bonding.
46. Hayes, S. A., Jones, F. R., Marshiya, K. & Zhang, W. A self-healing thermosetting composite material. *Composites A* **38**, 1116–1120 (2007).
47. Denissen, W., Winne, J. M. & Du Prez, F. E. Vitrimers: permanent organic networks with glass-like fluidity. *Chem. Sci.* **7**, 30–38 (2016).
48. Montarnal, D., Capelot, M., Tournilhac, F. & Leibler, L. Silica-like malleable materials from permanent organic networks. *Science* **334**, 965–968 (2011).
49. Scott, T. F., Schneider, A. D., Cook, W. D. & Bowman, C. N. Photoinduced plasticity in cross-linked polymers. *Science* **308**, 1615–1617 (2005).
50. Burnworth, M. *et al.* Optically healable supramolecular polymers. *Nature* **472**, 334–337 (2011).
51. Corten, C. C. & Urban, M. W. Repairing polymers using oscillating magnetic field. *Adv. Mater.* **21**, 5011–5015 (2009).
52. Cordier, P., Tournilhac, F., Soulié-Ziakovic, C. & Leibler, L. Self-healing and thermoreversible rubber from supramolecular assembly. *Nature* **451**, 977–980 (2008).
This paper describes the first thermoreversible self-healing rubber made by supramolecular assembly.
53. Hentschel, J., Kushner, A. M., Ziller, J. & Guan, Z. Self-healing supramolecular block copolymers. *Angew. Chem. Int. Edn Engl.* **51**, 10561–10565 (2012).
54. Das, A. *et al.* Ionic modification turns commercial rubber into a self-healing material. *ACS Appl. Mater. Interf.* **7**, 20623–20630 (2015).
55. Wang, Q. *et al.* High-water-content mouldable hydrogels by mixing clay and a dendritic molecular binder. *Nature* **463**, 339–343 (2010).
56. Lu, Y.-X., Tournilhac, F., Leibler, L. & Guan, Z. Making insoluble polymer networks malleable via olefin metathesis. *J. Am. Chem. Soc.* **134**, 8424–8427 (2012).
57. Imato, K. *et al.* Self-healing of chemical gels cross-linked by diarylbibenzofuranone-based trigger-free dynamic covalent bonds at room temperature. *Angew. Chem. Int. Edn Engl.* **51**, 1138–1142 (2012).
58. Yoon, J. A. *et al.* Self-healing polymer films based on thiol–disulfide exchange reactions and self-healing kinetics measured using atomic force microscopy. *Macromolecules* **45**, 142–149 (2012).
59. Dry, C. M. Procedures developed for repair of polymer matrix composite materials. *Compos. Struct.* **35**, 263–269 (1996).
60. White, S. R. *et al.* Autonomic healing of polymer composites. *Nature* **409**, 794–797 (2001).
This paper demonstrates the first autonomic self-healing of a fracture in a structural polymer by dispersed liquid-monomer-filled microcapsules and reactive catalyst particles.
61. Diesendruck, C. E., Sottos, N. R., Moore, J. S. & White, S. R. Biomimetic self-healing. *Angew. Chem. Int. Edn Engl.* **54**, 10428–10447 (2015).
62. Brown, E. N., White, S. R. & Sottos, N. R. Microcapsule induced toughening in a self-healing polymer composite. *J. Mater. Sci.* **39**, 1703–1710 (2004).
63. Jones, A. S., Rule, J. D., Moore, J. S., Sottos, N. R. & White, S. R. Life extension of self-healing polymers with rapidly growing fatigue cracks. *J. R. Soc. Interf.* **4**, 395–403 (2007).
64. Kang, S., Baginska, M., White, S. R. & Sottos, N. R. Core-shell polymeric microcapsules with superior thermal and solvent stability. *ACS Appl. Mater. Interf.* **7**, 10952–10956 (2015).
65. Kim, B., Jeon, T. Y., Oh, Y.-K. & Kim, S.-H. Microfluidic production of semipermeable microcapsules by polymerization-induced phase separation. *Langmuir* **31**, 6027–6034 (2015).
66. Norris, C. J. *et al.* Autonomous stimulus triggered self-healing in smart structural composites. *Smart Mater. Struct.* **21**, 094027 (2012).
67. Norris, C. J., Meadway, G. J., O'Sullivan, M. J., Bond, I. P. & Trask, R. S. Self-healing fibre reinforced composites via a bioinspired vasculature. *Adv. Funct. Mater.* **21**, 3624–3633 (2011).
68. Toohey, K. S., Sottos, N. R., Lewis, J. A., Moore, J. S. & White, S. R. Self-healing materials with microvascular networks. *Nature Mater.* **6**, 581–585 (2007).
This paper describes the self-healing of a polymer coating by monomer-filled, interconnected microvascular channels made by direct-write assembly of a sacrificial scaffold.
69. Hamilton, A. R., Sottos, N. R. & White, S. R. Self-healing of internal damage in synthetic vascular materials. *Adv. Mater.* **22**, 5159–5163 (2010).
70. Hansen, C. J. *et al.* Self-healing materials with interpenetrating microvascular networks. *Adv. Mater.* **21**, 4143–4147 (2009).
71. Hansen, C. J. *et al.* Accelerated self-healing via ternary interpenetrating microvascular networks. *Adv. Mater.* **21**, 4320–4326 (2011).
72. Huang, C.-Y., Trask, R. S. & Bond, I. P. Characterization and analysis of carbon fibre-reinforced polymer composite laminates with embedded circular vasculature. *J. R. Soc. Interf.* **7**, 1229–1241 (2010).
73. Trask, R. S. & Bond, I. P. Bioinspired engineering study of Plantae vasculature for self-healing composite structures. *J. R. Soc. Interf.* **7**, 921–931 (2010).
74. Theriault, D., White, S. R. & Lewis, J. A. Chaotic mixing in three-dimensional microvascular networks fabricated by direct-write assembly. *Nature Mater.* **2**, 265–271 (2003).
75. Patrick, J. F. *et al.* Continuous self-healing life cycle in vascularized structural composites. *Adv. Mater.* **26**, 4302–4308 (2014).
This paper describes the first repeated self-healing of a fibre-reinforced composite by two-part, reactive liquid delivery through 3D, interpenetrating microvascular networks.
76. Brochu, A. B. W., Chyan, W. J. & Reichert, W. M. Microencapsulation of 2-octylcyanoacrylate tissue adhesive for self-healing acrylic bone cement. *J. Biomed. Mater. Res. B* **100**, 1764–1772 (2012).
77. Brochu, A. B. W., Evans, G. A. & Reichert, W. M. Mechanical and cytotoxicity testing of acrylic bone cement embedded with microencapsulated 2-octylcyanoacrylate. *J. Biomed. Mater. Res. B* **102**, 181–189 (2014).
78. Gladman, A. S., Celestine, A.-D. N., Sottos, N. R. & White, S. R. Autonomic healing of acrylic bone cement. *Adv. Healthcare Mater.* **4**, 202–207 (2015).
79. Brochu, A. B. W., Craig, S. L. & Reichert, W. M. Self-healing biomaterials. *J. Biomed. Mater. Res. A* **96**, 492–506 (2011).
80. Odom, S. A. *et al.* Restoration of conductivity with TTF-TCNQ charge-transfer salts. *Adv. Funct. Mater.* **20**, 1721–1727 (2010).
81. Odom, S. A. *et al.* A Self-healing conductive ink. *Adv. Mater.* **24**, 2578–2581 (2012).
82. Odom, S. A. *et al.* Autonomic restoration of electrical conductivity using polymer-stabilized carbon nanotube and graphene microcapsules. *Appl. Phys. Lett.* **101**, 043106 (2012).
83. Blaiszik, B. J. *et al.* Autonomic restoration of electrical conductivity. *Adv. Mater.* **24**, 398–401 (2012).
84. Tee, B. C.-K., Wang, C., Allen, R. & Bao, Z. An electrically and mechanically self-healing composite with pressure- and flexion-sensitive properties for electronic skin applications. *Nature Nanotechnol.* **7**, 825–832 (2012).
85. Birnbaum, K. D. & Alvarado, A. S. Slicing across kingdoms: regeneration in plants and animals. *Cell* **132**, 697–710 (2008).
86. White, S. R. *et al.* Restoration of large damage volumes in polymers. *Science* **344**, 620–623 (2014).
This paper describes the self-healing of large-scale damage by vascular delivery of two-stage chemistry that first forms a dynamic gel scaffold and then polymerizes.
87. Kim, H., Mohapatra, H. & Phillips, S. T. Rapid, on-command debonding of stimuli-responsive crosslinked adhesives by continuous, sequential quinone methide elimination reactions. *Angew. Chem. Int. Edn Engl.* **54**, 13063–13067 (2015).
88. Baker, M. S., Kim, H., Olah, M. G., Lewis, G. G. & Phillips, S. T. Depolymerizable poly(benzyl ether)-based materials for selective room temperature recycling. *Green Chem.* **17**, 4541–4545 (2015).
89. Hernandez, H. L. *et al.* Triggered transience of metastable poly(phthalaldehyde) for transient electronics. *Adv. Mater.* **26**, 7637–7642 (2014).
90. Park, C. W. *et al.* Thermally triggered degradation of transient electronic devices. *Adv. Mater.* **27**, 3783–3788 (2015).
91. Kang, S.-K. *et al.* Bioresorbable silicon electronic sensors for the brain. *Nature* **530**, 71–76 (2016).
92. Diesendruck, C. E. *et al.* Mechanically triggered heterolytic unzipping of a low-ceiling-temperature polymer. *Nature Chem.* **6**, 623–628 (2014).
93. Jin, H. *et al.* Thermally stable autonomic healing in epoxy using a dual-microcapsule system. *Adv. Mater.* **26**, 282–287 (2014).
94. Maiti, S., Shankar, C., Geubelle, P. H. & Kieffer, J. Continuum and molecular-level modeling of fatigue crack retardation in self-healing polymers. *J. Eng. Mater. Technol.* **128**, 595–602 (2006).
95. Balazs, A. C. Modeling self-healing materials. *Mater. Today* **10**, 18–23 (2007).
96. Jones, A. S. & Dutta, H. Fatigue life modeling of self-healing polymer systems. *Mech. Mater.* **42**, 481–490 (2010).
97. Zhang, M. Q. & Rong, M. Z. Theoretical consideration and modeling of self-healing polymers. *J. Polym. Sci. B* **50**, 229–241 (2012).
98. Soghrati, S. *et al.* Computational analysis of actively-cooled 3D woven microvascular composites using a stabilized interface-enriched generalized finite element method. *Int. J. Heat Mass Transfer* **65**, 153–164 (2013).
99. Bluhm, J., Specht, S. & Schröder, J. Modeling of self-healing effects in polymeric composites. *Arch. Appl. Mech.* **85**, 1469–1481 (2015).
100. Nie, Z., Xu, S., Seo, M., Lewis, P. C. & Kumacheva, E. Polymer particles with various shapes and morphologies produced in continuous microfluidic reactors. *J. Am. Chem. Soc.* **127**, 8058–8063 (2005).
101. Tumbleston, J. R. *et al.* Continuous liquid interface production of 3D objects. *Science* **347**, 1349–1352 (2015).
102. Gladman, A. S., Matsumoto, E. A., Nuzzo, R. G., Mahadevan, L. & Lewis, J. A. Biomimetic 4D printing. *Nature Mater.* **15**, 413–418 (2016).

Acknowledgements J.F.P. and M.J.R. are grateful to the Arnold and Mabel Beckman Foundation for financial support through the Beckman Institute Postdoctoral Fellowship Program. The authors thank the Air Force Office of Scientific Research for support through the Center of Excellence in Self-Healing, Regeneration, and Structural Remodeling, the National Science Foundation (DMR 1307354), the Defense Advanced Research Project Agency (FA8650-13-C-7347) and the BP International Center for Advanced Materials (ICAM). We are grateful to D. Loudermilk, C. Klinger, A. Jerez and I. Patrick for assistance with graphics, and G. Wilson and M. Andersson for insightful discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests, see go.nature.com/2gnh9ma. Readers are welcome to comment on the online version of this paper at go.nature.com/2gnh9ma. Correspondence should be addressed to S.R.W. (swhite@illinois.edu).

Printing soft matter in three dimensions

Ryan L. Truby^{1,2} & Jennifer A. Lewis^{1,2}

Light- and ink-based three-dimensional (3D) printing methods allow the rapid design and fabrication of materials without the need for expensive tooling, dies or lithographic masks. They have led to an era of manufacturing in which computers can control the fabrication of soft matter that has tunable mechanical, electrical and other functional properties. The expanding range of printable materials, coupled with the ability to programmably control their composition and architecture across various length scales, is driving innovation in myriad applications. This is illustrated by examples of biologically inspired composites, shape-morphing systems, soft sensors and robotics that only additive manufacturing can produce.

Additive manufacturing, which encompasses a broad range of light- and ink-based printing techniques that allow the digital design and fabrication of three-dimensional (3D) objects, is transforming the science and engineering of advanced materials. Unlike conventional manufacturing methods that require moulds, dies or lithographic masks, digital assembly makes it possible to rapidly turn computer-aided designs into complex 3D objects on demand. Several techniques have been introduced over the past four decades^{1–7} that use industrial and desktop 3D printers to pattern soft materials. So far, commercial 3D printers have focused mostly on rapid prototyping of 3D objects. Most printing platforms use soft materials in one of three forms: photocurable resins^{2,3}, polymer powders^{4,5} or thermoplastic monofilaments⁶.

To unleash the vast potential of additive manufacturing, new materials and printing methods are needed that enable fabrication involving different materials at high speeds and with high precision over large build volumes. The scientific impetus for this technology is the drive to create architected matter that has qualitatively new properties, but this requires unprecedented control over the material's composition, structure, function and dynamics. By providing the ability to make products on demand in both low production runs and with customized form factors (such as size and shape), additive manufacturing provides a strong economic driver for adoption across a range of industrial sectors, such as aerospace, automotive, biomedical, robotics, and much more. From the manufacturing of plastic air ducts in aircraft to customized orthodontics, orthotics and hearing-aid shells, 3D printing is beginning to disrupt conventional manufacturing and supply chains across the world⁸.

In this Review we describe soft matter and introduce the light- and ink-based 3D printing techniques that are used to pattern such materials, with an emphasis on enhancing feature resolution, printing speed and the integration of different materials. We then highlight several emerging applications, including biologically inspired architectures for structural applications, shape-morphing structures, soft sensors and robots. Discussion of the many advances in 3D-printed biomedical devices^{9,10}, human tissues^{11,12–17}, and optical¹⁸ and electronic devices^{19–22} are beyond the scope of this Review, but there are already several excellent reviews that cover these areas. Finally, we share our perspective on the future directions with the potential for greatest societal impact.

Defining soft matter

Soft matter encompasses a broad range of synthetic and biological materials, including thermoplastic, thermosetting and elastomeric

polymers, hydrogels, liquid crystals and granular media²³. These materials are composed of basic building blocks — polymer chains, molecules or particles — that can be easily moved and so allow deformation under shear or other external forces. During 3D printing, the constituents are solidified into 3D architectures with elastic moduli that span orders of magnitude, from squishy hydrogels (10–100 kPa)¹¹ to rigid epoxy composites (>10 GPa)²⁴. The viscoelasticity, compliance (ease of deformation) and toughness of the printed materials may also be tailored to enable them to readily undergo (and even recover from) large deformations.

Overview of 3D printing

In 3D printing, a computer-controlled translation stage typically moves a pattern-generating device, either in the form of laser optics or an ink-based printhead, to fabricate objects a layer at a time. During the printing process, patterned regions composed of resins, powders or inks are solidified to yield the desired 3D form. Simply put, these printed objects are tangible representations of the digital designs that guide the printing process. Since the inception of 3D printing, several basic printing techniques have been introduced (Fig. 1), enabling technological advances that range from rapid prototyping to the additive manufacturing of finished parts^{2–6}. The specific patterning and solidification process used by a given 3D printing method define the minimum feature size it can create (Fig. 2a) and the type of printable soft materials it can use (Fig. 2b–g). Variations on these basic methods have largely focused on improving printing resolution²⁵ and speed^{26–28}, and on integrating multiple materials in a given printed part^{15,29–35}.

Light-based 3D printing

The first 3D printing methods to emerge used light to sculpt objects through either the stereolithography (SLA) of photocurable resins^{2,3}, or the selective laser sintering (SLS) of polymeric powders⁵ (Fig. 1a,b). In SLA, a liquid resin is selectively photopolymerized by a rastering laser. Once a layer has been printed, a new layer of liquid resin is introduced and subsequently crosslinked in locally illuminated regions. This process is repeated, layer by layer, until the desired 3D object is complete. Newer methods, including digital projection lithography (DLP)^{26,36}, continuous liquid interface production (CLIP)²⁷ and two-photon polymerization (2PP)²⁵, are all based on this basic concept. However, unlike SLA, which relies on point-source illumination to pattern one volume element (a 'voxel') at a time, DLP and CLIP enable an entire

¹John A. Paulson School of Engineering and Applied Sciences, and ²Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts 02138, USA.

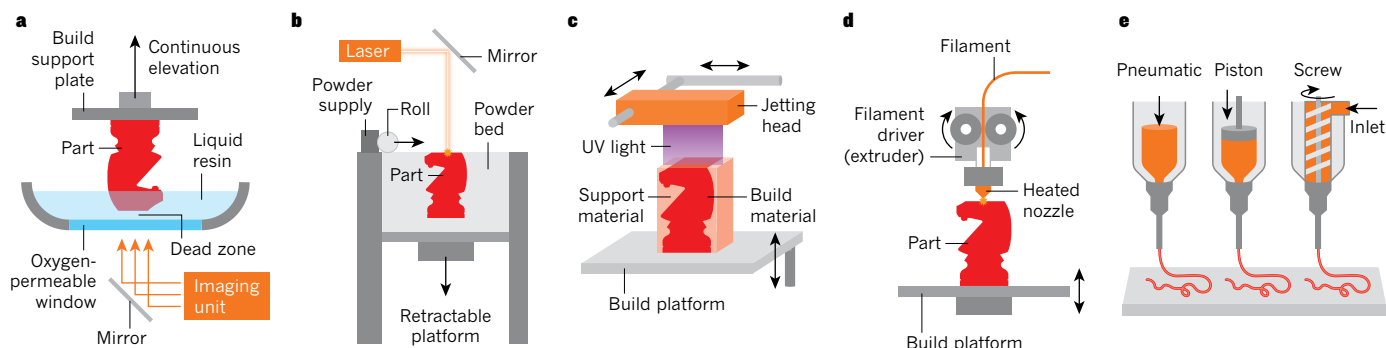


Figure 1 | Common light- and ink-based 3D printing methods. **a**, The light-based 3D printing method known as continuous liquid interface production. (Diagram adapted from ref. 27.) **b**, Light-based selective laser

layer to be solidified by using micro-mirror array devices²⁶ or dynamic liquid-crystal masks³⁶ to project a mask pattern onto the liquid-resin reservoir. As such, both DLP and CLIP are much faster than SLA. By contrast, 2PP provides the highest lateral resolution (around 100 nm) in 3D printed parts by taking advantage of the squared point-spread function associated with the two-photon absorption of light of wavelength λ , which is confined to a tightly focused voxel with dimensions on the order of λ^3 (ref. 25). But as with all 3D printing methods, there is an inherent trade-off between printer resolution (Fig. 2a), build volume and speed. This means that 2PP can be used to fabricate highly complex microarchitectures, but the overall dimensions are typically limited to 1 cm³ (Fig. 2b)^{25,37–39}, whereas CLIP can readily produce complex parts with overall dimensions exceeding 100 cm³ (Fig. 2c) with minimal surface roughness, but with lower lateral resolution²⁷. However, none of these methods currently allows multiple materials to be patterned in a single build sequence.

In SLS, polymer particles in a powder bed are locally heated and fused together by a rastering laser^{5,40}. After a layer has been printed, a new layer of powder is spread across the bed and locally sintered. To facilitate spreading, granulated powders are used that typically have diameters between 10 μ m and 100 μ m. The non-fused regions in the powder bed serve as a support material during the building process. After the 3D object has been completed and removed from the powder bed, the loose powder is removed and recycled⁵. A representative part produced by the SLS of nylon powder is shown in Fig. 2d. The minimum feature size achieved by this printing method is around 100 μ m, which is a few times larger than the typical particle size in the powder bed.

sintering of powders. **c**, Light- and ink-based photocurable inkjet printing of photopolymerizable resins. **d**, Ink-based fused deposition modelling of thermoplastic filaments. **e**, Direct ink writing using viscoelastic inks.

Ink-based 3D printing

Although light-based printing methods provide the highest feature resolution, they are limited to patterning with either photopolymerizable resins, which yield only rigid thermoset polymers, or thermoplastic polymer powders. Ink-based 3D printing methods, in contrast, can pattern myriad soft materials in the form of printable inks that are formulated from a wide range of molecular, polymeric or particulate species. These can be chosen to achieve the desired flow behaviour — characterized by the ink's viscosity, surface tension, shear yield stress, and shear elastic and loss moduli — required for either droplet- or filament-based printing.

In droplet-based printing methods, soft materials are deposited by printheads similar to those used in the printing of 2D documents. Several 3D printing methods use this approach, including direct inkjet printing⁴¹, hot-melt printing¹⁹ and inkjet printing on a powder bed⁴. Inks for these approaches are composed of low-viscosity fluids. For example, in hot-melt printing, wax-based inks are heated during droplet formation and then solidify on impact. Other inkjet printers combine ink- and light-based printing in one platform: photocurable resins, for example, are polymerized when they are printed by illumination with an ultraviolet light source (Figs 1c and 2e). In an alternative to depositing the component material itself, binder solutions can be jetted onto powder beds to locally fuse particles in a method akin to SLS printing^{4,19}. In all these ink-based printing approaches, drop formation depends on both the properties of the ink material and the printing parameters, including the ink's density (ρ), viscosity (μ), surface tension (γ) and characteristic droplet length (L , which in most cases is the drop diameter), as well as

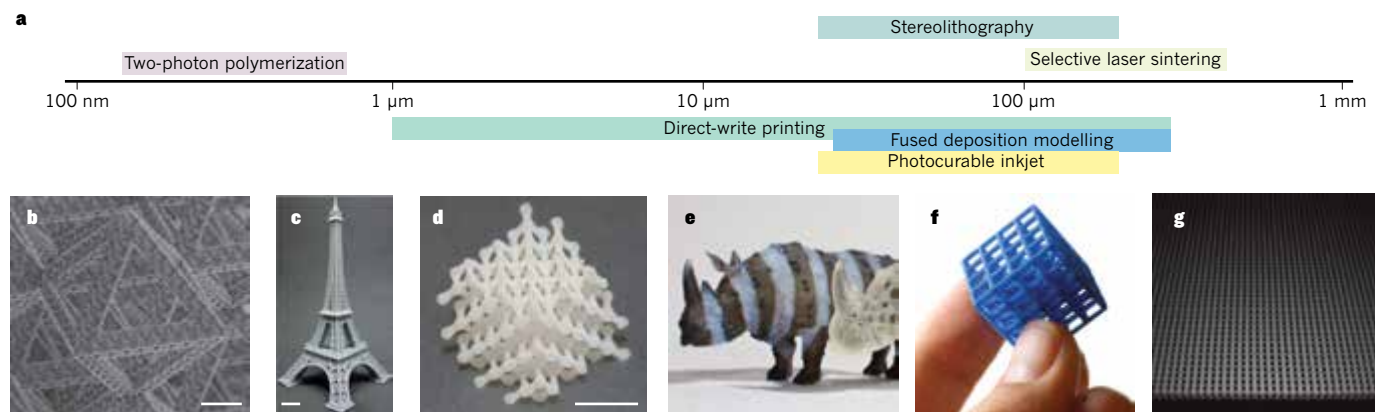


Figure 2 | Sizes and shapes of typical 3D-printed objects. **a**, Coloured bars show the minimum size ranges of patterned features produced by several light- and ink-based printing methods. **b–g**, Examples of polymer constructs printed by: **b**, two-photon polymerization (hierarchical octet truss; scale bar, 25 μ m; photo courtesy of J. Greer); **c**, continuous liquid interface production (Eiffel Tower; scale bar, 10 mm; adapted from ref. 27);

d, selective laser sintering (hierarchical lattice; scale bar, 10 mm; adapted from ref. 40); **e**, inkjet printing of photopolymerizable resins (multimaterial rhinoceros; adapted from ref. 54); **f**, fused deposition modelling (3D lattice; photo courtesy of S. Bernier, Zortrax); **g**, direct ink writing (3D epoxy lattice with 250- μ m features; photo courtesy of B. Compton and J. Lewis).

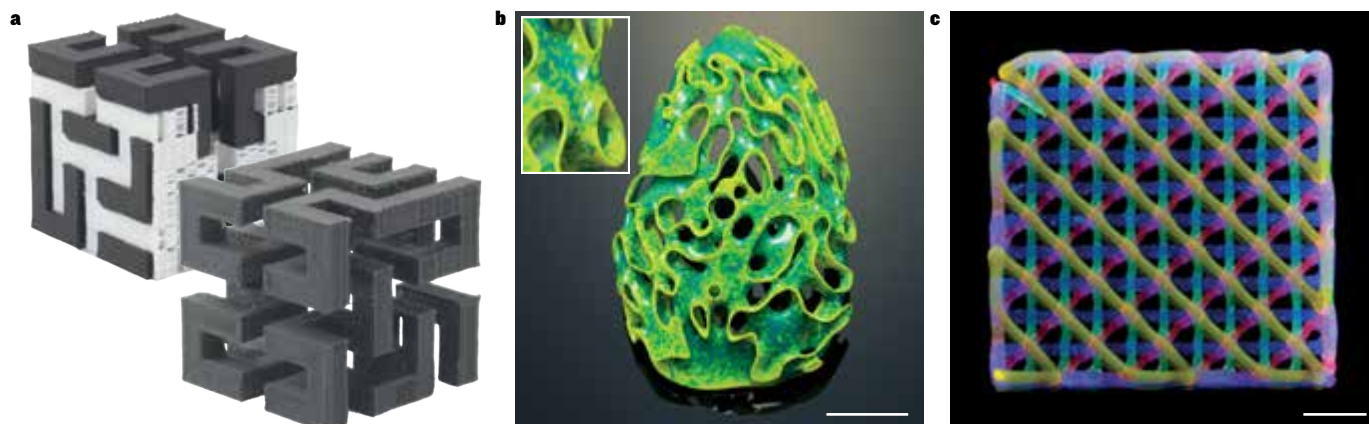


Figure 3 | Techniques for the fabrication of complex structures. **a**, The white sacrificial support material in an FDM-printed part (back) is removed to reveal a Hilbert cube with numerous overhangs (front). (Photo courtesy of Polymaker.) **b**, A conceptual artwork by N. Oxman produced by multimaterial inkjet printing (scale bar, 10 cm). The inset shows the complex distribution

of materials in the structure. (Photo courtesy of N. Oxman.) **c**, Multimaterial elastomeric lattice produced by direct ink writing. The 3D microlattice (1 cm × 1 cm × 1 mm) is produced by sequential printing layers composed of silicone-based inks dyed with blue, red, green and yellow fluorophores, each deposited by a separate nozzle (scale bar, 2 mm; adapted from ref. 15).

the velocity of the ejected droplet (v) and the nozzle diameter (d). These parameters must all be tightly controlled to achieve the right balance between viscosity, surface tension and inertial forces. This is usually captured by the dimensionless Z parameter, given as the inverse of the Ohnesorge number (Oh), that relates inertial and surface-tension forces to viscous forces as follows:

$$Z = 1/Oh = Re/\sqrt{We} = [\sqrt{(\rho\gamma L)}]/\mu \quad (1)$$

where Re and We are the Reynolds and Weber numbers, respectively^{41,42}. If viscous forces dominate (low Z), the ink droplets will not form during printing. If inertial or surface-tension forces dominate (high Z), ejected droplets will be prone to splashing or breaking up into multiple satellite droplets during printing, so print fidelity will diminish. Generally, ideal droplet formation occurs when Z is between 1 and 10, and the droplet velocity is at least equal to $\sqrt{(4\gamma/\rho d)}$. The fluid dynamics involved in drop formation, wetting and spreading play an important, yet limiting, role in defining the surface roughness and minimum feature size (~ 10 – 100 μm) of the printed objects. Typical values for μ , L and v are 2–20 mPa s, 10–30 μm and 1–10 m s^{-1} , respectively. All this means that it is difficult to jet (without clogging) complex fluids, such as concentrated polymer solutions, or solutions that contain filler particles that exceed 100 nm in diameter, or at concentrations above a few per cent. Nevertheless, these difficulties are in many cases outweighed by the huge advantages of inkjet-based methods arising from their highly sophisticated printhead designs — state-of-the-art multinozzle arrays may have thousands of nozzles that can deliver more than 100 million drops per second with picolitre volumes — and their ability to print using different materials⁴¹.

Compared with droplet-based methods, 3D filament printing allows a broader range of ink designs, feature sizes and geometries^{6,43}. In this approach, soft materials are deposited as a continuous filament but still a layer at a time. In the earliest form, known as fused deposition modelling (FDM), thermoplastic filaments are fed through a hot extrusion head during printing and then solidify as they cool below their glass transition temperature^{6,44} (Fig. 1d). Several types of thermoplastic polymer can be patterned by this approach, including the widely used acrylonitrile butadiene styrene (ABS), polylactic acid (PLA) and polycarbonate (Fig. 2f). The polymer filaments can also be filled with particles, such as carbon black, to enhance the functionality of the printed parts⁴⁵. Given their ease of use and compatibility with common materials, desktop FDM printers have helped to drive the ‘maker revolution’ in the past decade.

One important alternative to FDM printing is the direct ink writing (DIW) of viscoelastic materials under ambient conditions⁴³ (Fig. 1e).

Crucial to its success has been the development of concentrated polymer^{46–49}, fugitive organic (used as sacrificial materials)^{50,51}, and filled epoxy²⁴ inks, which have fluid properties that enable the printing of complex 3D architectures (Fig. 2g). These yield-stress fluids are well described by the Herschel–Bulkley model⁵²:

$$\tau = \tau_y + K\dot{\gamma}^n \quad (2)$$

where τ is the shear stress, K is the consistency $\dot{\gamma}$ is the shear rate, and n is the flow index ($n < 1$ for shear-thinning fluids). Typical values for the apparent ink viscosity, minimum filament diameter and printing speed are 10^2 – 10^6 mPa s (depending on the shear rate), 1–250 μm (~ 10 – 100 times higher than the characteristic size of the building blocks for a given ink), and 1 mm s^{-1} to 10 cm s^{-1} , respectively. To induce flow through the nozzle, the applied stress in the printhead must exceed the yield stress, τ_y , of these inks so that they fluidize and then, when they exit the nozzle, rapidly recover their original values of τ_y and the shear elastic modulus, G' (ref. 43).

In some cases, additional processing steps (such as photopolymerization or thermal curing) may be required to fully solidify the printed parts. When these steps are decoupled from the printing process, it can be difficult to build truly 3D objects, as the underlying printed layers may not fully support the subsequent layers. However, these problems can be overcome by using printheads coupled with ultraviolet LEDs⁵³ or heated build chambers.

Multimaterial 3D printing

The complexity and functional performance of 3D printed objects can be enhanced by printing different materials together, but this requires a high degree of spatial and compositional precision. Light-based methods are currently not well suited to such multimaterial fabrication, because it is difficult to dynamically alter the composition of a liquid photopolymer reservoir or powder bed during printing^{29,30}. By contrast, ink-based printing methods such as FDM, inkjet printing and DIW can easily be used for multimaterial 3D printing.

Both FDM and inkjet printers are capable of printing primary building materials alongside sacrificial materials that support overhanging or spanning features. An exemplary Hilbert cube produced by FDM is shown in Fig. 3a before and after the removal of the white support material. Inkjet printing enables voxel-by-voxel patterning of multiple materials, using a full-colour palette and photopolymer resins whose backbone composition, side-group chemistry and crosslink density can be systematically varied to produce regions with different mechanical properties (Fig. 3b), at a higher resolution than FDM printing can achieve^{35,54}.

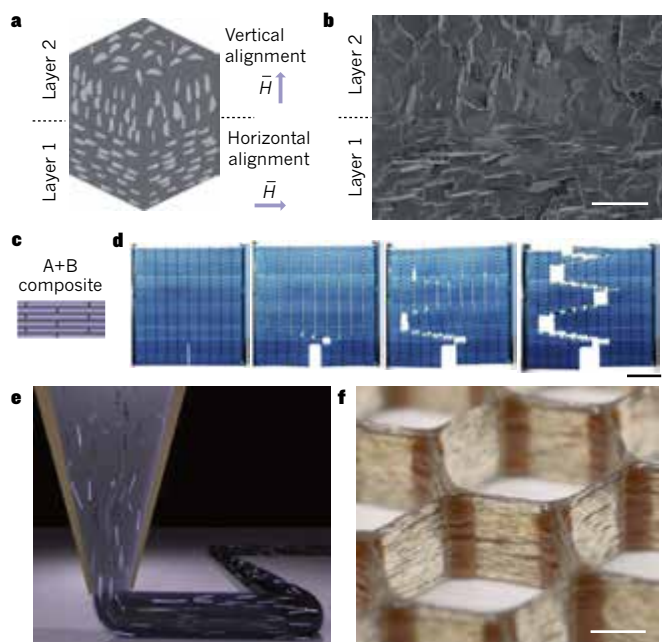


Figure 4 | Bio-inspired composites. **a, b**, 3D magnetic printing of platelet-reinforced composites, in which a magnetic field is used to induce the desired platelet orientation, and digital light projection (DLP) is used to locally photopolymerize oriented voxels (**a**). This motif mimics the layered architecture of abalone shells (**b**; scale bar, 25 μm ; both **a** and **b** are adapted from ref. 63). **c, d**, Inkjet printing of rigid (A) and compliant (B) material in a 'bricks and mortar' structure that resembles nacre or bone (**c**). Toughening occurs owing to delocalized load transfer away from the crack tip and crack deflection (**d**; scale bar, 20 mm; adapted from ref. 64). **e, f**, Direct ink writing of fibre-filled epoxy composites in a cellular motif inspired by balsa wood. The anisotropic fibre filler aligns in the shear and extensional flow field in the tapered nozzle during printing (**e**). An epoxy-based composite with hexagonal cells, in which carbon fibres align along the printing direction, that is, horizontally in the cell walls (**f**; scale bar, 2 mm; adapted from ref. 24).

At present, DIW⁴³ offers the broadest spectrum of printable materials, including structural^{11,47,48}, electrical^{33,55} and biological¹⁵ materials. Multimaterial DIW can be achieved either by using multiple (single-nozzle) printheads (Fig. 3c), each of which houses a different ink composition¹⁵, or by using microfluidic printheads that allow for switching³¹, mixing³², core-shell printing³³, or printing multiple filament arrays in a single pass²⁸. Microfluidic switching nozzles can swap between two different inks when required³¹, whereas mixing nozzles can be used to print materials with tunable gradients of mechanical, conductive or other material properties³². Core-shell printheads yield filaments that contain concentric layered materials³³. Finally, multinozzle printheads separate a single ink stream into 2^n streams, where n is the number of bifurcating generations in the printhead, allowing a dramatic reduction in build time (for example, a part requiring 24 h to build using a single nozzle can be printed in 22 min using a 64-nozzle array)²⁸. By using dual multinozzle arrays, two disparate inks can be patterned simultaneously. However, these multinozzle arrays consist of nozzles that are relatively large (100–200 μm in diameter), and they are not individually addressable like those used in inkjet printing. Finally, there is growing interest in directly writing inks into matrix materials by a process known as embedded 3D printing, which enables truly free-form fabrication of soft materials^{51,55,56}. These variants of DIW offer considerable flexibility in the types and motifs of shapes that can be printed.

Architected soft matter

The term 'architecture', which normally refers to the design and construction of buildings, is increasingly being used to describe materials that have optimized composition and topology. With 3D printing, it has become possible to fabricate architected matter from

an ever-broadening palette of soft materials in a programmable way, opening up a new design space for scientists and engineers^{8,16,19,21,57–59}. There are many noteworthy examples, but here we are focusing on advances in printing biologically inspired composites, shape-morphing systems, soft sensors and robotics.

Bio-inspired composites

Natural composite materials, such as nacre⁶⁰, bone⁶¹ and wood⁶², are typically held together by the organization of platelet or fibre reinforcement in complex architectures. These features help them achieve remarkable properties that exceed the sum of their parts, often combining stiffness, low density and high specific strength. They may also have energy-dissipation capabilities that lead to graceful failure, so they remain functional even when they start to fail. Inspired by these natural examples, researchers have focused on printing synthetic analogues in which the spatial organization and alignment of reinforcing fillers or printed features within polymer matrices are well controlled.

In one promising approach, external magnetic fields are used to control platelet orientation^{34,63} in photopolymerizable liquid resins, which are patterned layer-by-layer using DLP (Fig. 4a,b). The printer is modified by placing three electromagnetic solenoids around its periphery, which generate a magnetic field that aligns iron oxide-coated platelets (about 10 μm in length) suspended in the liquid photopolymer resin, along a prescribed vector in 3D space. The oriented voxels, whose minimum lateral dimension is about 100 μm , are photopolymerized to lock in the desired platelet orientation by crosslinking the surrounding matrix. Tensile testing reveals that printed objects with oxide platelets aligned parallel to the applied load exhibit higher stiffness (+29%), hardness (+23%) and strain at rupture (+100%) than those with orthogonally aligned platelets, and are twice as stiff as printed polymer matrices devoid of platelets. By coupling dynamic masking with magnetic alignment, filler particles can adopt different orientations within or between each layer (Fig. 4a). One architecture mimics the calcite prismatic and aragonite 'bricks and mortar' layers found in abalone shells⁶³ (Fig. 4b). There are limitations, however, owing to the sedimentation of dense fillers in the liquid resin during printing, which can lead to unintended compositional gradients, and excluded volume effects may hinder the orientation of filler in more concentrated systems.

Another approach to creating bricks-and-mortar architectures relies on multimaterial inkjet printing of rigid and compliant photocurable resins⁶⁴ (Fig. 4c). Samples composed entirely of either rigid (material A) or compliant (material B) material — the 'bricks' and 'mortar', respectively, in Fig. 4c — were printed, cured and characterized. Their respective yield strengths were 0.5 and 15 MPa, with a stiffness ratio, E_A/E_B , of about 1,500. In both cases, cracks initiate in the notched regions and propagate smoothly through the pure samples. Bio-inspired composites were also fabricated by printing rigid bricks coated with a thin compliant layer (about 250 μm thick). These architectures emulate the fracture-propagating, high-toughness properties of nacre and bone (Fig. 4d). Both delocalized load transfer away from the crack tip and crack deflection through the compliant coating enhance the fracture toughness of these printed composites. However, a little mixing (3–4%) occurs between layers during the printing process, reducing the effective stiffness ratio by nearly two orders of magnitude⁶⁴. To improve performance further, resin chemistries with more disparate baseline properties are needed to retain good interlayer adhesion during printing.

Some structural applications use fibre-filled epoxy composites in which the reinforcing fillers are in either discrete or continuous form. Inspired by balsa wood — which rivals the best engineering materials in terms of specific bending stiffness and strength — synthetic cellular architectures have been created by DIW using an epoxy resin-based ink filled with short carbon fibres. During the printing process, these anisotropic fillers align under the shear and extensional flow field that develops in the nozzle (Fig. 4e), resulting in enhanced

stiffness in the thermally cured composite along the printing direction (Fig. 4f). Printed tensile bars containing fibres aligned parallel to the applied load exhibited stiffness values nearly equivalent to those of wood cell walls, and 10–20 times higher than most commercial 3D-printed polymers²⁴. One shortcoming of DIW is its inability to fabricate continuous fibre-reinforced composites, but this is possible using a variant of FDM in which continuous fibres are embedded in thermoplastic matrices⁶⁵.

The patterned features and complexity of 3D-printed architectures do not yet match those found in nature. But there is scope to extend these boundaries and create materials with properties that meet or even exceed those of biological materials. If new materials and printing methods were capable of encoding a richer range of compositional and structural hierarchy across length scales, especially around 100 nm, this would accelerate innovation.

Shape-morphing systems

There is a growing emphasis on designing soft matter that has intrinsically programmed responsiveness, adaptability and other functionality. Materials of this sort include structural metamaterials, such as lightweight, ultra-stiff cellular trusses^{37,66}, topology-optimized auxetic⁴⁷ and negative-stiffness lattices⁶⁷, and bistable structures that store energy through mechanical deformation⁴⁸. A related and currently active research direction focuses on materials that autonomously change their shape. The term ‘4D printing’ is often used to describe the fabrication of 3D objects that can then change their shape over time in response to an environmental stimulus. Such shape-morphing systems often respond autonomously to light, heat or moisture, and are sometimes used in smart textiles⁶⁸, robotic systems⁶⁹ and biomedical devices⁷⁰.

In one approach, inkjet printing was used to create shape-changing architectures by patterning a light-absorbing ink onto a prestrained polystyrene substrate. Under infrared illumination, the underlying substrate was locally heated in the patterned regions, which acted like hinges to induce an autonomous, origami-like shape change⁷¹ (Fig. 5a,b). Building on this concept, linear structures with hinges that can swell have been created by multimaterial printing. These can self-assemble into various predetermined 3D shapes when immersed in water^{72,73} (Fig. 5c). In another approach, shape-memory polymers have been printed to create stimuli-responsive architectures^{74–77}. These constructs are fabricated in their intended (final) form before being warmed to a temperature above the glass transition temperature (T_g) of the hinges. They are then mechanically deformed to a prefolded or other initial shape, and cooled below T_g to lock the hinges in place. Upon reheating the printed object above T_g , its shape transforms back to the originally printed form^{74–77} (Fig. 5d). So far, only simple shape changes have been demonstrated.

Biomimetic 4D printing offers an easy route to encoding complex shape changes in hydrogel-based composites⁴⁹. Inspired by the nastic movements of plants^{78,79}, in which plants respond non-directionally to changes in stimuli such as heat or humidity, hydrogel inks containing stiff cellulose fibrils were designed to mimic the composition of plant cell walls. These anisotropic fibrils align along the printing direction, so it is possible to define the swelling and elastic anisotropies required to induce the desired shape change upon immersion in water by controlling the print path. Printing bilayer patterns in floral forms composed of five petals in either a 90°/0° or –45°/45° configuration can induce simple changes in curvature, such as bending and twisting, respectively, when the initially flat forms swell in water (Fig. 5e,f). A theoretical framework developed to solve the inverse problem (in which one wants to design a final form but the required print path is unknown) makes it possible to move beyond these simple structures to print much more complex shape-morphing architectures, including some that mimic orchids and calla lilies⁴⁹. The modularity of the composite inks used to fabricate these structures should make it possible to incorporate other hydrogel matrix and anisotropic filler chemistries to encode

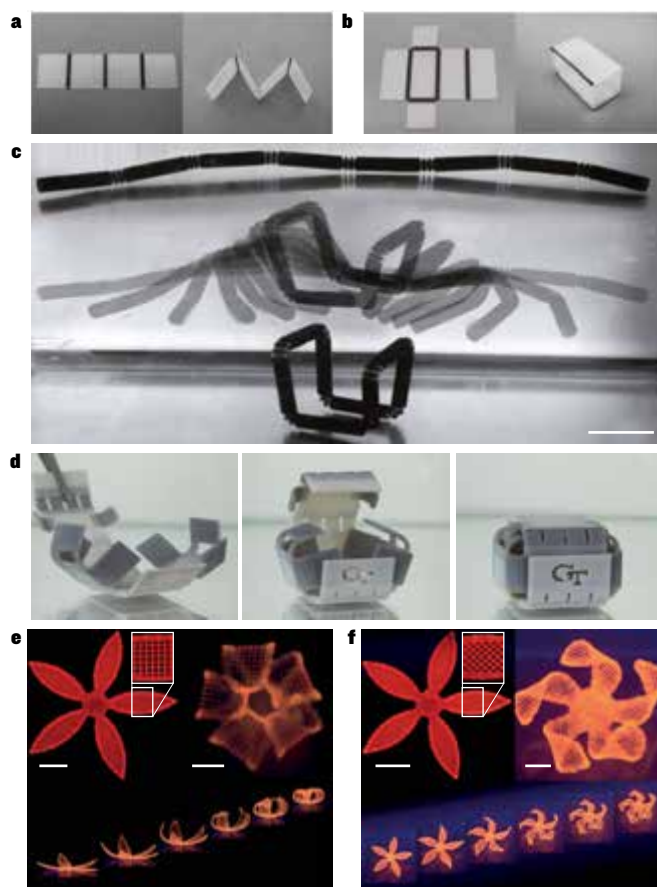


Figure 5 | Stimuli-responsive, morphing architectures. **a, b**, Prestrained polystyrene substrate with inkjet-printed hinges made of carbon black ink (**a**), which autonomously folds into a 3D shape (**b**) when illuminated with infrared light (scale bars, 10 mm; adapted from ref. 71). **c**, 4D-printed composite with swellable hinges (top) that self-assembles from a linear into a box-like structure (bottom) when immersed in water (scale bar, 5 cm; adapted from ref. 72). **d**, A 4D-printed unfolded box composed of shape-memory polymers that folds back into its original conformation when immersed in warm water (adapted from ref. 76). **e, f**, Biomimetic 4D printing of hydrogel composites containing anisotropic cellulose fibrils that orient along the printing direction. They undergo anisotropic swelling to programmably change shape when immersed in water. The printed bilayer lattices transform into flowers, whose petals either bend or twist when the bilayer orientations are 90°/0° (**e**) or –45°/45° (**f**) (scale bars, 5 mm; insets, 2.5 mm; adapted from ref. 49).

responses to other stimuli, such as light, heat and pH.

The focus is now turning to strategies for creating shape-morphing architectures that transform rapidly and provide significant actuation forces. However, the response times of shape-morphing structures are usually slow, and the structures tend to be mechanically weak — limitations that will need to be overcome if practical applications are to be developed.

Soft sensors and robots

Soft sensors, actuators and robots are improving human–machine interactions across a broad spectrum of applications. A central requirement for this is the ability to integrate soft materials with disparate mechanical and electrical properties in customized form factors^{80–82}; 3D printing is particularly well suited to produce such soft devices and systems.

Consider, for example, soft strain sensors, which are typically composed of a deformable conducting material that is patterned onto, attached to or encapsulated within an insulating, conformable, stretchable soft matrix^{21,83–85}. Embedded 3D printing has recently been used to fabricate highly stretchable strain sensors composed of a conductive carbon ink patterned in an elastomeric matrix⁵⁵ (Fig. 6a).

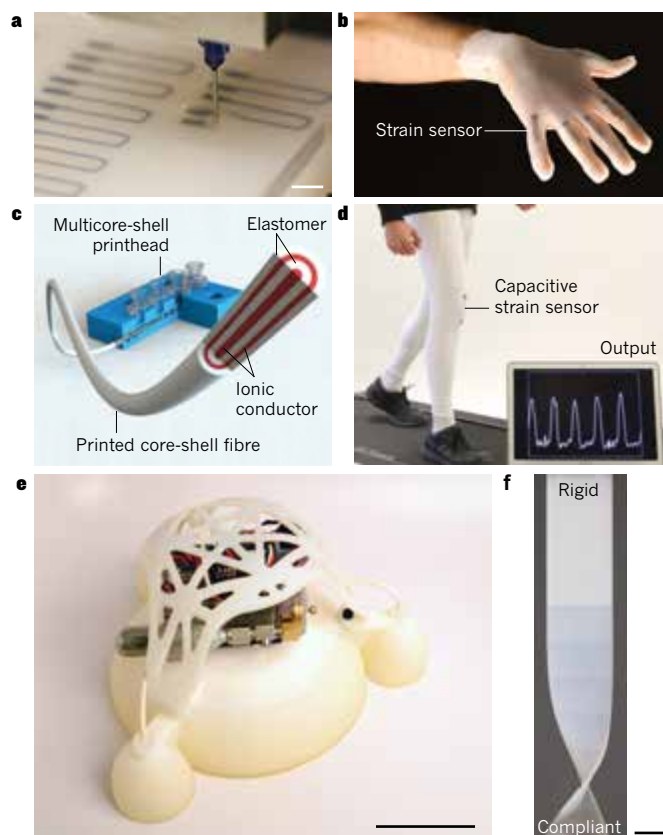


Figure 6 | Soft sensors, actuators and robots. **a, b**, Soft strain sensors are patterned directly in a free-form nature in an elastomeric matrix by ‘embedded 3D’ printing (**a**; scale bar, 5 mm). Strain sensors embedded into a glove-shaped, elastomeric matrix enable proprioceptive sensing of joint bending (**b**); (**a** and **b** adapted from ref. 55). **c, d**, Capacitive soft sensors based on ionically conductive inks are printed using a multicore-shell printhead, which produces a fibre sensor composed of concentrically layered materials. Concentric shells of ionically conductive ink (red) are encapsulated by dielectric, elastomer layers (white) (**c**). Soft capacitive sensors printed with multicore-shell printheads can be integrated into textiles for wearable technologies (**d**); (**c** and **d** adapted from ref. 33). **e, f**, A soft-bodied robot powered by combustion can carry heavy hardware (**e**; scale bar, 10 cm). The printed body has a graded modulus, enabling the compliant materials needed for locomotion to interface seamlessly with the rigid materials of the auxiliary hardware (**f**; scale bar, 10 mm); (**e** and **f** adapted from ref. 35).

The resulting hairpin sensors exhibit increased electrical resistance when they are stretched. The approach was then used to fabricate a wearable glove containing embedded strain sensors (Fig. 6b), which provide resistive feedback when the fingers are bent, making them ideal for training and rehabilitation purposes. The free-form nature of the embedded 3D printing allows the rapid fabrication of highly complex soft sensors⁵⁵, and avoids the delamination issues that typically arise for soft sensors made by conventional moulding and lamination processes^{86–88}. Other approaches for printing soft sensors have relied on directly printing elastomers, such as fluorinated rubbers with conductive particle fillers^{89,90}. One drawback of these sensors, however, is that they exhibit hysteresis — there is a time lag between the break-up and the reformation of the conductive particle networks during a given strain cycle^{55,89}.

The limitations caused by hysteresis can be overcome by integrating liquid metal (such as eutectic gallium indium, eGaIn) into soft sensing architectures^{86–88}. However, the high surface tension of eGaIn and other liquid metals poses serious challenges for printing⁹¹, so ionically conductive inks are being explored. These have recently been successfully encapsulated in a highly extensible elastomeric matrix and used to produce textile-mounted, capacitive fibre sensors³³. This

required a specially designed multicore-shell printhead that was capable of printing filaments composed of concentric conductive features separated by highly stretchable elastomer shells (Fig. 6c). The capacitance, resistance and decay time of these capacitive fibre sensors were measured as a function of strain, as required for soft joint proprioceptive sensing³³ (Fig. 6d).

Soft actuators derived from swellable hydrogels^{92–94}, granular media⁹⁵ and electroactive polymers⁹⁶ have all been fabricated so far. Of these, the most widely used are fluidic elastomer actuators (FEAs), which consist of a network of open channels within elastomeric composites^{80,82,97}. These embedded pneumatic networks inflate when filled with a fluid, inducing the desired actuating motion. Although FEAs are typically fabricated by a multistep moulding process, SLA printing of silicone-based photo-crosslinkable resins has recently been demonstrated. Using this method, FEAs can be designed with arbitrarily complex fluidic chambers to drive multidirectional actuation when inflated⁹⁸. Methods based on DIW have produced elastomeric actuators that serve as simple haptic feedback devices⁹⁹, and more complex FEAs have been produced by multimaterial inkjet printing¹⁰⁰. These initial demonstrations reveal the power of digital design and manufacturing, but further research is required to develop compatible materials systems, printing methods and predictive models to optimize soft actuator mechanics.

A final point regarding the soft robotic systems developed so far is that most require tethers to ancillary hardware for control and power. The interfacing of bulky, rigid hardware components with robots constructed from soft materials is not straightforward, however. Here, 3D printing can come into its own, as illustrated by the recent example of a soft robot that can jump being powered by combustion (Fig. 6e). The body of the robot was created using multimaterial inkjet printing to pattern multiple photopolymer layers of varying compliance. The resulting graded elastic modulus (Fig. 6f) meant that the robot body smoothly transitioned from a rigid core to a soft exterior, improving the interface between the robot’s body and the on-board power and control hardware needed for propulsion³⁵. Coupling 3D printing to appropriate design in this way offers tremendous opportunities for integrating soft control, power and sensing elements to create fully autonomous soft robots and machines¹⁰¹.

Future directions

Together, digital design and additive manufacturing have huge potential. The pace of discovery and innovation is rapidly accelerating as 3D, and now 4D, printing methods are increasingly embraced by the research community, as well as by industrial designers and engineers around the world.

From a scientific viewpoint, the ability to heterogeneously integrate soft materials with disparate mechanical, electrical and optical properties in topology-optimized architectures will lead to as-yet-unimagined performance. The examples highlighted above underscore the power of digital fabrication, but they should be viewed merely as a starting point. The current level of integration and sophistication in 3D-printed soft architectures is relatively simplistic; far more can be achieved by augmenting computer-aided design software with more informed inputs, perhaps based on materials genomics, multiscale modelling and topology optimization. But to fully take advantage of advanced generative designs, new 3D-printing platforms are also needed, so material composition and function can be controlled and designs integrated from the nanoscale to the macroscale. Closed-loop feedback control, coupled with machine vision and learning, would allow real-time error correction to ensure that 3D-printed objects conform to the target designs in a reproducible manner.

From a technological viewpoint, the adoption of 3D printing is being driven by applications that benefit from customization and have small production runs. Yet all the initial applications, such as patient-specific orthodontics, rely solely on the ability to create complex 3D shapes, often from a single material. The true power of

digital manufacturing will be realized only when form and function are fully integrated. If 'complexity' is inherently free in 3D-printed objects — that is, if it is as simple to print a cube as it is to print an architected form such as a miniature Eiffel Tower — then the ability to embed function is also necessarily free. It merely requires the integration of different materials across multiple length scales that give rise to unprecedented properties.

The rapidly changing digital landscape already pervades our lives and affects the way we communicate, connect and share information. But when will digital manufacturing cross the divide from niche applications to widespread adoption? This transition is already under way, as can be seen in the rapid growth in the use of desktop 3D printers by educators, makers and entrepreneurs, and the growing installation of more-sophisticated 3D printers for industrial manufacturing. Yet digital fabrication is hindered by several factors, including long build times, high cost and poor scalability. Moreover, most 3D printers have been developed for rapid prototyping, not manufacturing. For 3D printing to transform high-throughput manufacturing, either large numbers of low-cost desktop printers need to be deployed whose capabilities will improve over time, or new 3D printers must be developed that enable the continuous production of parts at high speeds. Either way, the convergence of advanced materials, hardware and software is inevitable, and these must be mastered in the twenty-first century. ■

Received 30 May; accepted 15 August 2016.

1. Swainson, W. K. Method, medium and apparatus for producing three-dimensional figure product. US patent 4,041,476 (1977).
2. Kodama, H. Automatic method for fabricating a three-dimensional plastic model with photo-hardening polymer. *Rev. Sci. Instrum.* **52**, 1770–1773 (1981).
3. Hull, C. W. Apparatus for production of three-dimensional objects by stereolithography. US patent 4,575,330 (1986).
4. Sachs, E. M., Haggerty, J. S., Cima, M. J. & Williams, P. A. Three-dimensional printing techniques. US patent 5,205,055 (1993).
5. Beaman, J. J. & Deckard, C. R. Selective laser sintering with assisted powder handling. US patent 4,938,816 (1990).
6. Crump, S. S. Apparatus and method for creating three-dimensional objects. US patent 5,121,329 (1992).
7. Bradshaw, S., Bowyer, A. & Haufe, P. The intellectual property implications of low-cost 3D printing. *ScriptEd* **7**, 5–31 (2010).
8. Lipson, H. & Kurman, M. *Fabricated: The New World of 3D Printing* (John Wiley, 2013).
9. Morrison, R. J. *et al.* Mitigation of tracheobronchomalacia with 3D-printed personalized medical devices in pediatric patients. *Sci. Transl. Med.* **7**, 285ra64 (2015).
10. Gupta, M. K. *et al.* 3D printed programmable release capsules. *Nano Lett.* **15**, 5321–5329 (2015).
11. Malda, J. *et al.* 25th anniversary article: Engineering hydrogels for biofabrication. *Adv. Mater.* **25**, 5011–5028 (2013).
12. Derby, B. Printing and prototyping of tissues and scaffolds. *Science* **338**, 921–926 (2012).
13. Villar, G., Graham, A. D. & Bayley, H. A Tissue-like printed material. *Science* **340**, 48–52 (2013).
14. Mannoor, M. S. *et al.* 3D printed bionic ears. *Nano Lett.* **13**, 2634–2639 (2013).
15. Kolesky, D. B. *et al.* 3D bioprinting of vascularized, heterogeneous cell-laden tissue constructs. *Adv. Mater.* **26**, 3124–3130 (2014).
16. Murphy, S. V. & Atala, A. 3D bioprinting of tissues and organs. *Nature Biotechnol.* **32**, 773–785 (2014).
17. Ma, X. *et al.* Deterministically patterned biomimetic human iPSC-derived hepatic model via rapid 3D bioprinting. *Proc. Natl Acad. Sci. USA* **113**, 2206–2211 (2016).
18. Gissibl, T., Thiele, S., Herkommer, A. & Giessen, H. Two-photon direct laser writing of ultracompact multi-lens objectives. *Nature Photonics* **10**, 554–560 (2016).
19. de Gans, B.-J., Duineveld, P. C. & Schubert, U. S. Inkjet printing of polymers: state of the art and future developments. *Adv. Mater.* **16**, 203–213 (2004).
20. Kong, Y. L. *et al.* 3D printed quantum dot light-emitting diodes. *Nano Lett.* **14**, 7017–7023 (2014).
21. Rim, Y. S., Bae, S. H., Chen, H., De Marco, N. & Yang, Y. Recent progress in materials and devices toward printable and flexible sensors. *Adv. Mater.* **28**, 4415–4440 (2016).
22. Kong, Y. L., Gupta, M. K., Johnson, B. N. & McAlpine, M. C. 3D printed bionic nanodevices. *Nano Today* **11**, 330–350 (2016).
23. Jones, R. A. L. *Soft Condensed Matter* (Oxford Univ. Press, 2002).
24. Compton, B. G. & Lewis, J. A. 3D-printing of lightweight cellular composites. *Adv. Mater.* **26**, 5930–5935 (2014).
25. Cumpston, B. H. *et al.* Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. *Nature* **398**, 51–54 (1999).
26. Sun, C., Fang, N., Wu, D. M. & Zhang, X. Projection micro-stereolithography using digital micro-mirror dynamic mask. *Sensors Actuators A* **121**, 113–120 (2005).
27. Tumbleston, J. R. *et al.* Continuous liquid interface production of 3D objects. *Science* **347**, 1349–1352 (2015).
28. Hansen, C. J. *et al.* High-throughput printing via microvascular multinozzle arrays. *Adv. Mater.* **25**, 96–102 (2013).
29. Choi, J.-W., MacDonald, E. & Wicker, R. Multi-material microstereolithography. *Int. J. Adv. Manuf. Technol.* **49**, 543–551 (2010).
30. Choi, J.-W., Kim, H.-C. & Wicker, R. Multi-material stereolithography. *J. Mater. Process. Technol.* **211**, 318–328 (2011).
31. Hardin, J. O., Ober, T. J., Valentine, A. D. & Lewis, J. A. Microfluidic printheads for multimaterial 3D printing of viscoelastic inks. *Adv. Mater.* **27**, 3279–3284 (2015).
32. Ober, T. J., Foresti, D. & Lewis, J. A. Active mixing of complex fluids at the microscale. *Proc. Natl Acad. Sci. USA* **112**, 12293–12298 (2015).
33. Frutiger, A. *et al.* Capacitive soft strain sensors via multicore-shell fiber printing. *Adv. Mater.* **27**, 2440–2446 (2015).
34. Kokkinis, D., Schaffner, M. & Studart, A. R. Multimaterial magnetically assisted 3D printing of composite materials. *Nature Commun.* **6**, 8643 (2015).
35. Bartlett, N. W. *et al.* Robot powered by combustion. *Science* **349**, 161–165 (2015).
36. Zheng, X. *et al.* Design and optimization of a light-emitting diode projection micro-stereolithography three-dimensional manufacturing system. *Rev. Sci. Instrum.* **83**, 125001 (2012).
37. Zheng, X. *et al.* Ultralight, ultrastiff mechanical metamaterials. *Science* **344**, 1373–1377 (2014).
38. Meza, L. R. *et al.* Resilient 3D hierarchical architected metamaterials. *Proc. Natl Acad. Sci. USA* **112**, 11502–11507 (2015).
39. Frenzel, T., Findeisen, C., Kadic, M., Gumbsch, P. & Wegener, M. Tailored buckling microlattices as reusable light-weight shock absorbers. *Adv. Mater.* **28**, 5865–5870 (2016).
40. Kinstlinger, I. S. *et al.* Open-source selective laser sintering (OpenSLS) of nylon and biocompatible polycaprolactone. *PLoS ONE* **11**, e0147399 (2016).
41. Derby, B. Inkjet printing of functional and structural materials: fluid property requirements, feature stability, and resolution. *Annu. Rev. Mater. Res.* **40**, 395–414 (2010).
42. Fromm, J. E. Numerical calculation of the fluid dynamics of drop-on-demand jets. *IBM J. Res. Dev.* **28**, 322–333 (1984).
43. Lewis, J. A. Direct ink writing of 3D functional materials. *Adv. Funct. Mater.* **16**, 2193–2204 (2006).
44. Zein, I., Hutmacher, D. W., Tan, K. C. & Teoh, S. H. Fused deposition modeling of novel scaffold architectures for tissue engineering applications. *Biomaterials* **23**, 1169–1185 (2002).
45. Farahani, R. D., Dubé, M. & Theriault, D. Three-dimensional printing of multifunctional nanocomposites: Manufacturing techniques and applications. *Adv. Mater.* **28**, 5794–5821 (2016).
46. Gratson, G. M., Xu, M. & Lewis, J. A. Direct writing of three-dimensional webs. *Nature* **428**, 386 (2004).
47. Clausen, A., Wang, F., Jensen, J. S., Sigmund, O. & Lewis, J. A. Topology optimized architectures with programmable Poisson's ratio over large deformations. *Adv. Mater.* **27**, 5523–5527 (2015).
48. Shan, S. *et al.* Multistable architected materials for trapping elastic strain energy. *Adv. Mater.* **27**, 4296–4301 (2015).
49. Gladman, A. S., Matsumoto, E. A., Nuzzo, R. G., Mahadevan, L. & Lewis, J. A. Biomimetic 4D printing. *Nature Mater.* **15**, 413–418 (2016).
50. Theriault, D., Shepherd, R. F., White, S. R. & Lewis, J. A. Fugitive inks for direct-write assembly of three-dimensional microvascular networks. *Adv. Mater.* **17**, 395–399 (2005).
51. Wu, W., Deconinck, A. & Lewis, J. A. Omnidirectional printing of 3D microvascular networks. *Adv. Mater.* **23**, H178–H183 (2011).
52. Herschel, W. H. & Bulkley, R. Konsistenzmessungen von Gummi-Benzollösungen. *Kolloid Z.* **39**, 291–300 (1926).
53. Farahani, R. D., Lebel, L. L. & Theriault, D. Processing parameters investigation for the fabrication of self-supported and freeform polymeric microstructures using ultraviolet-assisted three-dimensional printing. *J. Micromech. Microeng.* **24**, 055020 (2014).
54. Vidimce, K., Wang, S.-P., Ragan-Kelley, J. & Matusik, W. OpenFab: A programmable pipeline for multi-material fabrication. *ACM Trans. Graph.* **32**, 136 (2013).
55. Muth, J. T. *et al.* Embedded 3D printing of strain sensors within highly stretchable elastomers. *Adv. Mater.* **26**, 6307–6312 (2014).
56. Bhattacharjee, T. *et al.* Writing in the granular gel medium. *Sci. Adv.* **1**, e1500655 (2015).
57. Brackett, D., Ashcroft, I. & Hague, R. Topology optimization for additive manufacturing. In *Solid Freeform Fabrication Symposium* 348–362 (2011).
58. Lin, D. *et al.* Three-dimensional printing of complex structures: Man made or

- toward nature? *ACS Nano* **8**, 9710–9715 (2014).
59. Montemayor, L., Chernow, V. & Greer, J. R. Materials by design: Using architecture in material design to reach new property spaces. *MRS Bull.* **40**, 1122–1129 (2015).
 60. Wegst, U. G. K., Bai, H., Saiz, E., Tomsia, A. P. & Ritchie, R. O. Bioinspired structural materials. *Nature Mater.* **14**, 23–36 (2015).
 61. Launey, M. E., Buehler, M. J. & Ritchie, R. O. On the mechanistic origins of toughness in bone. *Annu. Rev. Mater. Res.* **40**, 25–53 (2010).
 62. Gibson, L. J. The hierarchical structure and mechanics of plant materials. *J. R. Soc. Interface* **9**, 2749–2766 (2012).
 63. Martin, J. J., Fiore, B. E. & Erb, R. M. Designing bioinspired composite reinforcement architectures via 3D magnetic printing. *Nature Commun.* **6**, 8641 (2015). doi:10.1038/ncomms9641
 - This paper demonstrates the power of combining 3D printing with external magnetic fields to produce oriented composite architectures.**
 64. Dimas, L. S., Bratzel, G. H., Eylon, I. & Buehler, M. J. Tough composites inspired by mineralized natural materials: Computation, 3D printing, and testing. *Adv. Funct. Mater.* **23**, 4629–4638 (2013).
 65. Matsuzaki, R. *et al.* Three-dimensional printing of continuous-fiber composites by in-nozzle impregnation. *Sci. Rep.* **6**, 23058 (2016).
 66. Schaedler, T. A. *et al.* Ultralight metallic microlattices. *Science* **334**, 962–965 (2011).
 67. Duoss, E. B. *et al.* Three-dimensional printing of elastomeric, cellular architectures with negative stiffness. *Adv. Funct. Mater.* **24**, 4905–4913 (2014).
 68. Hu, J., Meng, H., Li, G. & Ibekwe, S. I. A review of stimuli-responsive polymers for smart textile applications. *Smart Mater. Struct.* **21**, 053001 (2012).
 69. Felton, S., Tolley, M., Demaine, E., Rus, D. & Wood, R. A method for building self-folding machines. *Science* **345**, 644–646 (2014).
 70. Randall, C. L., Gultepe, E. & Gracias, D. H. Self-folding devices and materials for biomedical applications. *Trends Biotechnol.* **30**, 138–146 (2012).
 71. Liu, Y., Boyles, J. K., Genzer, J. & Dickey, M. D. Self-folding of polymer sheets using local light absorption. *Soft Matter* **8**, 1764–1769 (2012).
 72. Tibbitts, S. 4D printing: Multi-material shape change. *Architect. Des.* **84**, 116–121 (2014).
 - This paper describes the first embodiment of 4D printing.**
 73. Raviv, D. *et al.* Active printed materials for complex self-evolving deformations. *Sci. Rep.* **4**, 7422 (2014).
 74. Ge, Q., Qi, H. J. & Dunn, M. L. Active materials by four-dimension printing. *Appl. Phys. Lett.* **103**, 131901 (2013).
 75. Ge, Q., Dunn, C. K., Qi, H. J. & Dunn, M. L. Active origami by 4D printing. *Smart Mater. Struct.* **23**, 094007 (2014).
 76. Mao, Y. *et al.* Sequential self-folding structures by 3D printed digital shape memory polymers. *Sci. Rep.* **5**, 13616 (2015).
 77. Mao, Y. *et al.* 3D printed reversible shape changing components with stimuli responsive materials. *Sci. Rep.* **6**, 24761 (2016).
 78. Burgert, I. & Fratzl, P. Actuation systems in plants as prototypes for bioinspired devices. *Phil. Trans. R. Soc. A* **367**, 1541–1557 (2009).
 79. Guo, Q. *et al.* Fast nastic motion of plants and bioinspired structures. *J. R. Soc. Interface* **12**, 20150598 (2015).
 80. Ilievski, F., Mazzeo, A. D., Shepherd, R. F., Chen, X. & Whitesides, G. M. Soft robotics for chemists. *Angew. Chem. Int. Edn Engl.* **50**, 1890–1895 (2011).
 81. Bauer, S. *et al.* 25th anniversary article: A soft future: From robots and sensor skin to energy harvesters. *Adv. Mater.* **26**, 149–162 (2014).
 82. Rus, D. & Tolley, M. T. Design, fabrication and control of soft robots. *Nature* **521**, 467–475 (2015).
 83. Yamada, T. *et al.* A stretchable carbon nanotube strain sensor for human-motion detection. *Nature Nanotechnol.* **6**, 296–301 (2011).
 84. Lu, N., Lu, C., Yang, S. & Rogers, J. Highly sensitive skin-mountable strain gauges based entirely on elastomers. *Adv. Funct. Mater.* **22**, 4044–4050 (2012).
 85. Lee, C., Jug, L. & Meng, E. High strain biocompatible polydimethylsiloxane-based conductive graphene and multiwalled carbon nanotube nanocomposite strain sensors. *Appl. Phys. Lett.* **102**, 183511 (2013).
 86. Majidi, C., Kramer, R. & Wood, R. J. A non-differential elastomer curvature sensor for softer-than-skin electronics. *Smart Mater. Struct.* **20**, 105017 (2011).
 87. Park, Y.-L., Chen, B. & Wood, R. J. Design and fabrication of soft artificial skin using embedded microchannels and liquid conductors. *IEEE Sens. J.* **12**, 2711–2718 (2012).
 88. Chossat, J. B., Park, Y.-L., Wood, R. J. & Duchaine, V. A soft strain sensor based on ionic and metal liquids. *IEEE Sens. J.* **13**, 3405–3414 (2013).
 89. Sekitani, T. *et al.* A rubberlike stretchable active matrix using elastic conductors. *Science* **321**, 1468–1472 (2008).
 90. Matsuhashi, N. *et al.* Printable elastic conductors with a high conductivity for electronic textile applications. *Nature Commun.* **6**, 7461 (2015).
 91. Boley, J. W., White, E. L., Chiu, G. T. C. & Kramer, R. K. Direct writing of gallium-indium alloy for stretchable electronics. *Adv. Funct. Mater.* **24**, 3501–3507 (2014).
 92. Lee, H., Xia, C. & Fang, N. X. First jump of microgel: actuation speed enhancement by elastic instability. *Soft Matter* **6**, 4342–4345 (2010).
 93. Palleau, E., Morales, D., Dickey, M. D. & Velev, O. D. Reversible patterning and actuation of hydrogels by electrically assisted ionoprinting. *Nature Commun.* **4**, 2257 (2013).
 94. Ionov, L. Biomimetic hydrogel-based actuating systems. *Adv. Funct. Mater.* **23**, 4555–4570 (2013).
 95. Brown, E. *et al.* Universal robotic gripper based on the jamming of granular material. *Proc. Natl Acad. Sci. USA* **107**, 18809–18814 (2010).
 96. Anderson, I. A., Gisby, T. A., McKay, T. G., O'Brien, B. M. & Calius, E. P. Multi-functional dielectric elastomer artificial muscles for soft and smart machines. *J. Appl. Phys.* **112**, 041101 (2012).
 97. Suzumori, K. Elastic materials producing compliant robots. *Rob. Auton. Syst.* **18**, 135–140 (1996).
 98. Peele, B. N., Wallin, T. J., Zhao, H. & Shepherd, R. F. 3D printing antagonistic systems of artificial muscle using projection stereolithography. *Bioinspir. Biomim.* **10**, 055003 (2015).
 99. Robinson, S. S. *et al.* Integrated soft sensors and elastomeric actuators for tactile machines with kinesthetic sense. *Extreme Mech. Lett.* **5**, 47–53 (2015).
 100. MacCurdy, R., Katzschnmann, R., Kim, Y. & Rus, D. Printable hydraulics: A method for fabricating robots by 3D co-printing solids and liquids. Preprint at <http://arxiv.org/abs/1512.03744> (2015).
 101. Wehner, M. *et al.* An integrated design and fabrication strategy for entirely soft, autonomous robots. *Nature* **536**, 451–455 (2016).

Acknowledgements We thank J. Raney, J. Muth and M. Skylar-Scott for their valuable insights, and the Wyss Institute for Biologically Inspired Engineering.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests, see go.nature.com/2gopvxx. Readers are welcome to comment on the online version of this paper at go.nature.com/2gopvxx. Correspondence and requests for materials should be addressed to J.A.L. (jalewis@seas.harvard.edu).

Reviewer information *Nature* thanks B. Derby and D. Theriault for their contributions to the peer review of this work.

The rise of plastic bioelectronics

Takao Someya^{1,2}, Zhenan Bao³ & George G. Malliaras⁴

Plastic bioelectronics is a research field that takes advantage of the inherent properties of polymers and soft organic electronics for applications at the interface of biology and electronics. The resulting electronic materials and devices are soft, stretchable and mechanically conformable, which are important qualities for interacting with biological systems in both wearable and implantable devices. Work is currently aimed at improving these devices with a view to making the electronic–biological interface as seamless as possible.

Almost all commercially available bioelectronic devices^{1–6} rely on silicon microelectronics, which is the workhorse for modern information infrastructure and technologies, including health-care and medical devices. Indeed, many current medical implants and devices such as pacemakers, electrocardiogram sensors and smart endoscopes rely on silicon microchips^{5,6}. Advances in the miniaturization of silicon microelectronics with nanometre-scale accuracy have reduced the size of these electronic modules, allowing them to be used for single-point health monitoring. This change has been made possible by the rigidity and mechanical stability of the inorganic materials used.

Creating the next generation of implantable or wearable electronics will require the introduction of new features, however, including mechanical flexibility (Box 1), large-area and facile processing of thin films, controlled biological properties, and mixed electronic and ionic conductivity. Mechanical flexibility is particularly important for device components that are in direct contact with certain areas of the skin or soft tissue to minimize the discomfort of worn or attached electronics. Regardless of whether the active devices are made of inorganic, organic or hybrid materials, the use of plastic films as substrates affords significant weight and thickness reductions while maintaining mechanical robustness and flexibility⁷. In contrast to silicon semiconductors, using inherently soft electronic materials that have a low Young's modulus to directly contact biological tissues can minimize adverse reactions, owing to the improved mechanical compliance between the tissue and the implanted device^{8,9}.

As well as providing favourable mechanical properties for interfacing with biological tissue, plastic electronics offer the potential for large-area, multimodal, multipoint sensing or stimulation on curvilinear surfaces^{7,10} (Fig. 1). Indeed, the use of organic semiconducting polymers has rapidly expanded from flexible displays^{11,12}, which have already been commercialized, to more advanced (and bidirectional) devices such as flexible, stretchable sensors — so-called artificial skins^{13,14}. The challenge in moving from flexible displays to sensing functions is to find a way of monitoring the complex, dynamic structures of biological organs over a large area with high spatial and temporal resolution. Flexible large-area organic circuits with an active-matrix design can already be used to reduce both power consumption and the amount of wiring involved relative to 'traditional' electronic devices^{13,14}.

Furthermore, the diversity and synthetic tunability of plastic materials are expected to allow features such as biodegradability¹⁵ and printability¹⁶, while maintaining the benefits associated with their

softness and flexibility. The stimulus responsiveness of plastics also affords natural conformability to three-dimensional (3D) surfaces and changes in shape, and allows on-demand self-repair¹⁷. The printability of polymers is another favourable attribute for cost-competitiveness and ease of customization⁷. Cost is always a major consideration when it comes to commercialization, but disposability is the most effective way of avoiding infections in hospitals, and that can be costly. Customization is particularly important in clinical applications, as it enables devices to be made to suit the needs of individual patients. Finally, mixed electronic and ionic transport in conducting polymers also allows coupling with ions in biological media, enabling low-impedance contacts for efficient electrical recording and stimulation^{18,19}. Ionic transport in polymers can also enable drug delivery through processes such as passive leaching or even electrophoretic transport²⁰.

This Review will discuss the latest progress in the use of soft electronic materials and their related devices in biological interfaces, and highlight future research directions and challenges that remain to be overcome. We emphasize recent work that harnesses properties that are unique to polymeric electronic materials, and consider the corresponding benefits to bioelectronics. We also briefly discuss synergies with high-performance inorganic electronic materials, which are complementary and can be used cooperatively for hybrid bioelectronics.

Developments in materials

The biological interface of organic electronics is a relatively recent development, but organic electronics have been intensively studied and developed over the past half-century. They have been used in commercial applications such as photoconductors in photocopying and laser printing, electrochromic films, anticorrosion and antistatic coatings based on conducting polymers, organic light-emitting diode (OLED) displays and lighting, organic photovoltaic cells (OPVs) and organic thin-film transistors (OTFTs)¹⁰. Some conducting polymers have been shown to achieve metallic transport behaviour^{21–24}, and charge-carrier mobilities of more than $10\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$, which rival that of poly-Si, have been reported for organic semiconductors^{25–27}. The progress made towards soft implantable and wearable devices relies not only on these advances in conducting and semiconducting polymers, but also on additional biomimetic properties, such as stretchability, self-healing and biodegradability (Fig. 2).

Stretchability is essential for comfort while wearing, for intimate attachment to curved surfaces and moving parts, and for the

¹Department of Electrical and Electronic Engineering, University of Tokyo, Tokyo 113-8656, Japan. ²Center for Emergent Matter Science (CEMS), Riken, Saitama 351-0198, Japan. ³Department of Chemical Engineering, Stanford University, Stanford, California 94305, USA. ⁴Department of Bioelectronics, Ecole Nationale Supérieure des Mines, CMP-EMSE, MOC, 13541 Gardanne, France.

BOX 1

Developing materials for soft interfaces

It is difficult to develop materials for soft interfaces because electronics and semiconductor devices are typically made of silicon and inorganic semiconductors, which are rigid (they have a high Young's modulus of about 100 GPa), whereas biological tissues have a much lower Young's modulus (from 10 GPa for bone to 1 kPa or less for brain tissue)^{102,103}. In an attempt to introduce mechanical flexibility into health-monitoring systems, components that use very thin silicon membranes and/or chips embedded in thin polymer films have been proposed and demonstrated^{98,104,105}. One example is electronic tattoos, in which a silicon microchip a few micrometres thick, which is both flexible and stretchable, is laminated directly on the skin¹⁰⁴. Similar flexible and stretchable devices that have inorganic membranes can be used in devices to be implanted in the brain, heart and other organs^{106,107}. It has been recognized that mechanical flexibility can be achieved by using thin membranes of silicon or other inorganic semiconductors. However, reducing the size of silicon vertically or laterally does not change the Young's modulus, and there will still be a large mismatch in the mechanical properties of inorganic materials and biological tissues.

Material	Young's modulus	Strain-to-break
Silicon	130 GPa	1%
Bone	~20 GPa	1%
Plastics	1 GPa	5%
Elastomer	0.01–10 MPa	50–4,000%
Gel	1–1,000 kPa	10–2,000%
Brain	<1 kPa	20%

maintenance of mechanical robustness. Indeed, strain tolerance of more than 80% is required for devices that are mounted on the knuckle, and more than 50% for those worn on the knee joint. A combination of organic devices on ultrathin plastic substrates and prestrained elastic substrates yields stable electrical properties under repeated strain in excess of 100% (refs 28–30). Plastic nanocomposite electronic materials are showing promising performance as stretchable conductors. For example, metal nanowires, metal nanoparticles and nanoflakes, carbon nanotubes, graphene and combinations of these have been incorporated in stretchable plastic materials to achieve both conductivity above 100 S cm^{-1} and high stretchability of up to 100% strain^{31–36}. Some have also been found to have a 'programmable' response, in which the nanomaterials exhibit nanoscale buckling after the first strain release. Subsequent stretching to the same initial strain level maintained about the same conductance, even after thousands of stretch–release cycles³¹. However, if rigid 'island' structures are connected with stretchable wires, even larger strain tolerance on the wires will be required than if inherently stretchable wiring or conductors were used. Some recently reported materials can maintain a conductivity above 100 S cm^{-1} , even at above 100% strain³². Conductivity values greater than 100 S cm^{-1} are sufficient for most practical sensors, but much higher conductivity is needed for neural stimulators. Plasticizers have been found to significantly reduce the elastic modulus of poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS) and significantly increase the stretchability^{37,38}, but the addition of a plasticizer reduces the conductivity of the resulting polymer. It will be important to find a way of maintaining the same conductance under different strain levels for polymer conductors.

For the semiconducting components of devices, regio-regular poly(3-hexylthiophene) (P3HT) and its block copolymer with polyethylene can

be plastically deformed to strains of more than 300% (ref. 39), although their charge-carrier mobilities are low, around $10^{-2} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ at only 0% strain; further significant irreversible decrease has been observed under strain. Some polymers have been found to exhibit higher mobilities and less reduction under strains of about 100%. However, the reversibility and strain–release cycling stability of organic semiconducting materials still need to be improved^{17,40,41}. Semiconducting carbon nanotubes and semiconducting polymer nanofibres have been shown to maintain charge-carrier mobilities and endure high strains of up to 100%, and semiconducting carbon nanotubes can also maintain high mobility^{34,42,43}.

Biodegradability and self-healing are also required if plastic bioelectronics are to have more-biomimetic properties. So far, the development of biodegradable plastic electronics has focused mainly on making devices on biodegradable substrates, regardless of the active materials. This is because the substrate constitutes more than 99% by weight (wt%) of the entire device, including sensors and electronic circuits. Biodegradable substrates that have commonly been used include aliphatic polyester-based biodegradable polymers, silk and cellulose^{44–47}. Several metal electrode materials have also been found to be biodegradable and biocompatible under certain conditions⁴⁸. These have been combined with biodegradable substrates and used in implantable medical devices^{48,49}. Additionally, thin silicon membranes have been found to give high-performance bio-resorbable electronics, providing new opportunities for bioelectronics^{50–52}. By contrast, only a limited number of biodegradable and biocompatible conducting and semiconducting organic materials have been reported so far. Attempts are being made to design and develop synthetic biodegradable conducting polymers. However, the conductivity values (currently at $10^{-4} \text{ S cm}^{-1}$) still need to be improved significantly⁵³.

Self-healing is essential for biological systems, and incorporating some form of autonomous and repeatable self-healing into electronic devices would enhance their robustness and durability, allowing them to be used in long-term implants and devices. But only a few studies have investigated self-healing in electronic devices, so there is an opportunity to make great improvements. Self-healing can be readily achieved by incorporating dynamic bonds in insulating polymer gels, such as hydrogen bonds, electrostatic interactions, and metal–ligand bonds⁵⁴. One study has reported a self-healing conducting polymer with conjugated cores crosslinked by reversible bonds between *N*-heterocyclic carbenes and transition metals⁵⁵, although the conductivity of the polymer is only around $10^{-3} \text{ S cm}^{-1}$. Composites of metal particles and self-healing polymer are the most likely candidates to achieve both high conductivity and autonomous repeated healing. There have been reports of the potential applications of such materials, such as electronic skin, transparent electrodes, and binders for battery electrodes^{48,49,54–58}.

Current applications

Two main areas for plastic bioelectronics are currently being pursued: wearable (non-invasive) devices and implantable devices.

Wearables and beyond

The super-conformability and stretchability of ultrathin-film plastic devices make them ideal for use in the next generation of wearables, which will be attached directly to the living, moving surface of human skin²⁸. Electrically, these materials have been demonstrated in electronic artificial skin (e-skin) with the use of organic transistors, for possible applications in robotics¹³. In this development, scalable circuits, which are designed for use in stretchable large-area sensors, use organic active matrices to measure pressure and temperature distributions¹³.

Regardless of where wearable electronics are attached, there are two features of plastic and organic electronic devices that make them particularly well suited for use in wearable devices: their excellent mechanical durability, and their potentially large area. Various plastic and organic electronic devices, such as OTFTs²⁸, OLEDs²⁹ and OPVs⁵⁹, have been fabricated on 1- μm -thick film substrates, which are just a

a Electronically functional polymers and/or organic electronics

Functions: biological, physical, mechanical, chemical and electronic

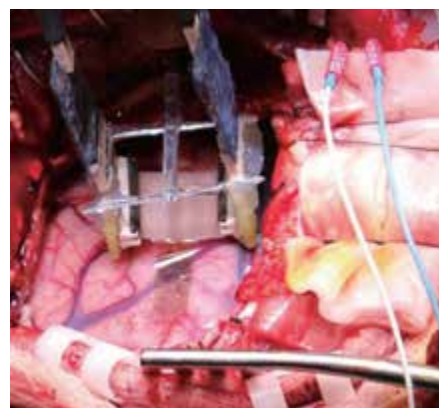
**b****E-skins and wearables****Implantable devices**

Figure 1 | The diversity of plastic bioelectronics. Electrically functional polymers and organic electronics provide a multifunctional, soft bio-interface. **a**, Polymers may have physical functions (thermal, acoustic and photonic), chemical functions (from surface modification and chemical interactions), mechanical functions (adhesiveness and softness), electronic functions (electric sensing and stimulation) and biological functions (biocompatibility).

b, Plastic bioelectronic devices have a range of applications. Left, a flexible sensor array that detects pressure can be laminated on a robot's hands as artificial skin (e-skin)^{13,14}. Reducing the thickness of sensor arrays to 1 μm allows devices to be ultraflexible, ultralightweight and stretchable, so they can be applied to the human body^{28,100}. Right, plastic bioelectronics can be implanted for neural recording, drug delivery and cell control, for example⁸³.

tenth of the thickness of kitchen wrap. Reducing the thickness of the substrate reduces the weight of the device and improves its bendability and conformability, because the strain induced by bending the film decreases proportionally as the thickness is decreased. These organic integrated circuits have been found to exhibit extraordinary robustness despite being super-thin — indeed, their electrical properties and mechanical performance were practically unchanged, and no degradation was observed, when they were squeezed to a bending radius of 5 μm , dipped in physiological saline, and stretched to up to double their original size.

To collate, compute and communicate the vast amount of data acquired by wearable sensors, flexible digital circuits such as processors⁶⁰, shift registers⁶¹ and memories⁶², as well as wireless circuits⁶³, have been developed. Although many state-of-the-art wearable devices are connected to rigid digital circuits, such flexible elements should be chosen appropriately and integrated with rigid, high-performance, inorganic semiconductor devices, so that the mechanical and electronic requirements may be satisfied simultaneously. In addition to digital and wireless circuits, analogue circuits, such as amplifiers, may also be required, because of the low magnitude of biological signals, which typically range from tens of microvolts in electroencephalography to millivolts in electrocardiography. To position the first-stage amplifier as close as possible to where the signals are generated, flexible amplifiers with a power gain exceeding 50 dB for a bandwidth beyond 1 kHz have been reported⁶⁴.

As well as semiconductor devices, there are various types of unique polymeric sensor. For bioelectronic applications, such sensors are broadly classified into two categories: physical sensors, which measure temperature, pressure, strain and light, for example; and (bio-)chemical sensors, such as ion, DNA, metabolite and protein sensors. Physical sensors are made from polymers to provide softness, which enables the measurement of pressure sensitivities of up to a few pascals^{65,66}. These

sensors are most sensitive at about body temperature^{67,68}. Conversely, (bio-)chemical sensors use the material diversity and synthetic flexibility of polymers to achieve greater specificity and sensitivity. Polymer transistors modified with odorant-binding proteins can provide sensitive and quantitative measurement of the weak interactions associated with neutral enantiomers⁶⁹, and allow for the sensitive and dynamic monitoring of cells for toxicology⁷⁰ without requiring reporter molecules. Chemical information such as oxygen concentration in the blood can be measured by using organic photonics comprising OLEDs and organic photodetectors (OPDs)^{71,72}.

Plastic integrated circuits and devices can be manufactured in large numbers by printing on large-area plastic films. Transistors with sub-micrometre channels have also been fabricated by surface modification and inkjet printing⁷³. Furthermore, a prototype of a wearable electronic circuit was recently printed on a 1- μm -thick film by exploiting the film's thinness and large area⁷⁴. In the age of the 'internet of things', the ability to customize wearable sensors will lead to an increase in 'on-demand' digital fabrication, and a combination of 3D and inkjet printing is likely to be crucial to meeting these needs.

Implantable devices

Implants traditionally rely on hard electronic materials, but these often elicit a 'foreign body' response, which limits their lifetime. This is a major limitation of neural implants, which have been developed for research purposes, for the diagnosis and treatment of various pathologies such as epilepsy and Parkinson's disease, and for brain-machine interfaces that seek to restore lost function. A typical example is the use of microfabricated silicon shuttles, which have metal electrodes that penetrate the brain and record neural activity. The use of soft organic coatings on the metal electrodes is being explored as a strategy for improving stability⁷⁵. Tuning the mechanical properties of these coatings leads to a variety of forms, including hydrogels that have

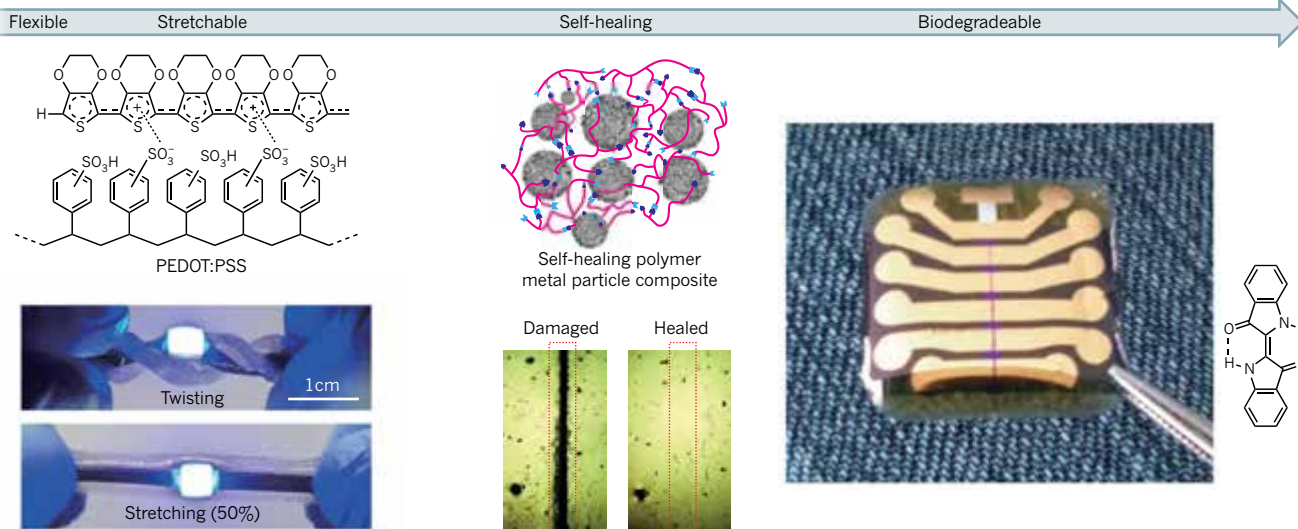


Figure 2 | Soft electronic polymers for plastic bioelectronics can be stretchable, biodegradable and have self-healing properties. Left, chemical structure of the conducting polymer PEDOT:PSS, which can be made stretchable by adding a surfactant³⁸. Reproduced with permission of Wiley-VCH Verlag from ref. 38. Middle, a self-healing conductive material made

from a self-healing polymer and nickel particles with nanospikes⁵⁸. Right, a sheet of organic field-effect transistors made with natural and biodegradable materials: shellac substrate, aluminium oxide and tetratetracontane dielectric, and indigo semiconductor. Electrodes were made of aluminium and gold. Reproduced with permission of Wiley-VCH Verlag from ref. 101.

mechanical properties similar to those of brain tissue⁷⁶. In some situations it is possible to harness the soft nature of polymer substrates to replace the hard shuttle altogether. Such implants can be inserted in the brain by using a temporary shuttle that is removed after insertion⁷⁷, or they can be placed on the brain or on a nerve where they can access neurons close to the surface. Using only polymer substrates can reduce the mismatch of mechanical properties at the biotic–abiotic interface, resulting in longer-lasting implants. This has been demonstrated by silicone-based spinal-cord implants with stretchable metal electrodes, which show excellent bio-integration with the central nervous system⁷⁸.

Using polymer coatings can improve the mechanical properties of devices, but it can also significantly lower the impedance of metal microelectrodes and enable high-quality recording and efficient electrical stimulation of neurons in laboratory animals^{79–81}. The uptake of ions from the biological environment into electronic polymers is exploited in organic electrochemical transistors for signal amplification. This improves the signal-to-noise ratio of brain recordings, as shown in animal-based models of epilepsy⁸². Most of the research has concentrated on using such coatings on hard, penetrating shuttles that help to access different areas in the brain, and combining them with soft, conformable polymeric substrates has been particularly effective for interfaces with the cortex. The use of PEDOT:PSS microelectrodes on thin poly(p-xylylene) film has provided single-neuron recordings from the surface of a rat's brain⁸³. Because these devices do not penetrate the brain, they are already being used on human patients diagnosed with epilepsy for high-resolution intraoperative recordings. More recently, PEDOT:PSS electrodes have been combined with flexible electronics and sensors on silicone elastomers that were cast and cured on 3D models of the epicardium. These hybrid devices showed improved electrical recording characteristics in animal models⁸⁴. Finally, transparent graphene electrodes integrated with poly(p-xylylene) substrates have been shown to enable the simultaneous use of various optical techniques including optogenetics, fluorescence microscopy, and 3D optical coherence tomography⁸⁵.

The flexible fabrication offered by organic materials has led to new ways of interacting with living systems. For example, *in situ* polymerization of conducting polymers in the brain is seen as a potential way of rebuilding the charge transport pathways across the glial scars caused by an implant. PEDOT that is grown in the hippocampus of rats does not seem to disable their memory, as observed by the way they navigate

a maze⁸⁶. Conducting polymers grown inside hydrogels and seeded with live cells are also being developed with the objective of creating 'living electrodes' that can establish new neural connections between an implanted device and the brain⁸⁷.

The delivery of drugs such as neurotrophins and anti-inflammatory molecules *in vivo* is being used to reduce the inflammatory response to a foreign-body implant, and more generally for controlled drug delivery past the blood–brain barrier. Polypyrrole-coated electrodes loaded with neurotrophin-3, for example, can be used for the simultaneous electrical and biochemical stimulation of cochlear neurons. Using a guinea-pig model, the release of neurotrophin-3 was shown to have beneficial effects on the auditory brainstem response threshold and on the density of the spiral ganglion neurons that survive implantation⁸⁸. In addition, a device called an organic electronic ion pump (OEIP) uses plastic electronics to achieve the dry electrophoretic delivery of ions from a reservoir to a target tissue. And OEIPs that deliver neurotransmitters have been used to tune the sense of hearing²⁰, reduce pain⁸⁹ in animal models, and stop epilepsy-like activity⁹⁰ in a brain-slice model.

Nerve regeneration and repair is another emerging application of plastic bioelectronics. This work is motivated by the *in vitro* demonstration that electrical stimulation through a conducting polymer can enhance the outgrowth of neurites⁹¹. *In vivo* electrical stimulation of sciatic-nerve defects in a rat model by using conducting polymer scaffolds has also been shown to promote axonal regeneration and remyelination⁹².

Many other devices have been tested *in vitro* and are being developed for use as implantable devices in the clinic. These include a variety of physical and biological sensors that can be used for multimodal sensing. For example, a conformable thermal sensor has been developed⁶⁷ that uses organic circuitry on a plastic substrate to resolve spatial temperature gradients on the surface of a lung. When combined with electrophysiology, such devices can provide valuable information about the functioning of the human body. Other examples include devices that use conducting polymers to electrically control cell adhesion⁹³ and signalling⁹⁴. Devices of this sort are potentially applicable to the diagnosis and treatment of diseases such as cancer, and the engineering of tissues for organ regeneration and replacement. Other examples include photoconducting, conjugated polymer-based layers, which show promise for the restoration of vision in explants of blind rat retinas⁹⁵. All these devices bring unique capabilities to the interface with biology that

go far beyond simple electrical recording and stimulation of neurons. Coupling them with soft polymeric substrates may deliver advanced 'multi-implants' that could one day potentially be inserted under the skin or be implanted deeper in the body through minimal openings, or even be injected by a syringe⁹⁶.

Challenges and prospects

The first tangible goal of plastic bioelectronics is the development of next-generation user interfaces for machines, and the second goal is advanced health care. With regard to the first goal, comfortable controls for prosthetic limbs and skeleton robot suits are needed to develop a system that can estimate the exact amount of force required to perform a task. And the accurate monitoring of sensations and emotions will have an important role in the creation of intelligent robots that can perceive human feelings and respond accordingly. In these applications, plastic electronics can be used to monitor and stimulate the skin, using a vast number of sensors. Direct control by a brain-machine interface could be possible if a large-area, high-density implantable plastic multiplexing system is used to connect electronics with neurons in the brain. Devices for medical applications will largely use the same platform as non-medical plastic devices, although the goals of the two types of device will differ. For healthcare devices, minimum invasiveness is required, but it is essential to maintain function and high performance.

Efforts aimed at implementing biologically inspired principles of operation will require systems with different architectures from conventional von Neumann systems, in which the physical separation between processing and memory limits throughput. Such systems would be adaptive, fault tolerant and would require little power, making them suitable for handling signals from a variety of biosensors. Indeed, electronic touch sensors have recently been used to transform the intensity of pressure signals to frequency-modulated spiky signals, which are characteristic of animal skin and nerve cells in general (including brain cells), and even to directly stimulate the brains of mice⁹⁷.

But several scientific and engineering challenges need to be overcome before we can fully exploit the benefits of plastic bioelectronics in practical devices. For a start, we currently have only a limited understanding of electronic-biological interfaces, so it will be important to have a theoretical model of complex systems that include water and ions. We also need a better understanding of the interplay of molecular design rules if we are to incorporate multiple functions of soft materials, such as charge transport, stretchability, degradation control and self-healing. Because plastic bioelectronics is a new and multidisciplinary field, it is expected to dovetail with other emerging fields, such as microfluidics⁹⁸ for drug delivery, and the study of induced pluripotent stem cells for regenerative therapy, for example.

From an engineering viewpoint, one of the biggest challenges facing plastic electronics — particularly plastic bioelectronics — is data analysis, because they generate large amounts of new types of data. Recently developed methods for handling huge amounts of data, and machine-learning technology, will be required for the analysis of the enormous amounts of data flowing in from the biosensors that are being deployed in this emerging field. Potential applications for bioelectronic devices, such as high-resolution neural recording of the brain, and 24-hour monitoring of metabolite and disease-marker concentrations in the blood, will generate complex data, which must be analysed to determine their biological meaning.

The long-term environmental stability and mechanical durability of plastic devices must be improved, and devices on the skin and other organs will need to be permeable to gases and moisture. Some bioelectronic devices require direct contact with aqueous media that contain large concentrations of salts, proteins and other biological molecules, and this must not affect their ability to function. So several questions remain about the long-term chemical and physical stability of exposed electronic surfaces and the effects of the body on their electronic and mechanical properties. Finding solutions will require the use of materials that are stable when exposed to air and water for the parts that make

contact with the biological environment, and encapsulation technologies are needed to protect the parts that do not. A plastic device by itself is mechanically durable, but there is a need to ensure the mechanical robustness of the entire system by establishing reliable electric interconnections between the soft elements (such as conductive gels and stretchable conductors) and the rigid elements (miniature batteries and silicon wireless chips).

The development of large-volume production facilities is also important for the creation of a new industry. In particular, handling ultrathin and rubbery substrates is a big challenge. Having disposable plastic sensors would substantially reduce the risk of infections in hospitals, especially for devices that directly touch the skin, but the production of such components must be cost-effective. Ultimately, high-throughput production lines need to be developed by combining roll-to-roll processes with digital fabrication, such as inkjet printing, to achieve self-alignment and fine resolution on plastic substrates, which are easily deformed. Once such production lines are established, printable inorganic materials such as carbon-nanotube inks and solution-processed polycrystalline silicon, as well as semiconducting polymers, can be used to further improve the electronic performance of large-area sensors⁹⁹.

Finally, non-technical issues will also have a bearing on future developments in plastic bioelectronics. The ethics of data collection, storage and analysis is a challenge facing products developed for the internet of things, especially for devices that regulate or monitor human health, regardless of whether they are based on plastic or other materials. Non-technical issues also have a major role in determining the commercial viability of any new biomedical or clinical use, especially for implantable devices. The biocompatibility of devices made from new materials requires strict evaluation, leading developers to be conservative in adopting new materials, especially in implanted devices. For this reason, the first clinical applications of plastic bioelectronics are likely to be *in vitro* diagnostics or cutaneous devices. The subsequent demonstration of significant gains in performance (for example, a lower-impedance conducting polymer coating that extends the battery life of a stimulator) and the enabling of new capabilities (such as a low-impedance coating that is capable of drug delivery) will provide strong incentives for implant manufacturers to adopt plastic bioelectronics, and accelerate support from doctors and patients.

The ultimate goal of plastic bioelectronics is the development of seamless, bidirectional interfaces between humans and machines. A huge number of challenges face materials and devices for large-area, multipoint and multimodal sensors on 3D curved, dynamically moving, living objects. But synergies between plastic or organic materials and high-performance inorganic materials for hybrid devices will accelerate and expand the development of bioelectronics. And one day it will seem normal to have a bionic interface and to interact with plastic bioelectronics as an integral part of the body. ■

Received 9 December 2015; accepted 1 September 2016.

1. Berggren, M. & Richter-Dahlfors, A. Organic bioelectronics. *Adv. Mater.* **19**, 3201–3213 (2007).
2. Rivnay, J., Owens, R. M. & Malliaras, G. G. The rise of organic bioelectronics. *Chem. Mater.* **26**, 679–685 (2014).
3. Wallace, G. G., Moulton, S. E. & Wang, C. *Proc. SPIE* **7642**, 764202 (2010).
4. Liao, C. *et al.* Flexible organic electronics in biology: materials and devices. *Adv. Mater.* **27**, 7493–7527 (2015).
5. Fitzpatrick, D. *Implantable Electronic Medical Devices* (Elsevier, 2015).
6. Lay-Ekuakille, A. & Mukhopadhyay, S. C. *Wearable and Autonomous Biomedical Devices and Systems for Smart Environment* (Springer, 2010).
7. Embracing the organics world. *Nature Mater.* **12**, 591 (2013).
8. Freed, L. E., Engelmayer, G. C., Borenstein, J. T., Moutos, F. T. & Guilak, F. Advanced material strategies for tissue engineering scaffolds. *Adv. Mater.* **21**, 3410–3418 (2009).
9. Delmas, P., Hao, J. & Rodat-Despoix, L. Molecular mechanisms of mechanotransduction in mammalian sensory neurons. *Nature Rev. Neurosci.* **12**, 139–153 (2011).
10. Bredas, J.-L. & Marder, S. R. *The WSPC Reference on Organic Electronics: Organic Semiconductors: Organic Semiconductors* (World Scientific Publishing, 2016).
11. Crawford, G. *Flexible Flat Panel Displays* (John Wiley, 2005).
12. Rogers, J. A. *et al.* Paper-like electronic displays: Large-area rubber-stamped plastic sheets of electronics and microencapsulated electrophoretic inks. *Proc.*

- Natl Acad. Sci. USA* **98**, 4835–4840 (2001).
13. Someya, T. *et al.* A large-area, flexible pressure sensor matrix with organic field-effect transistors for artificial skin applications. *Proc. Natl Acad. Sci. USA* **101**, 9966–9970 (2004).
This paper reports that an artificial electronic skin has been made using flexible multipoint sensors with an active matrix circuit.
 14. Someya, T. *et al.* Conformable, flexible, large-area networks of pressure and thermal sensors with organic transistor active matrixes. *Proc. Natl Acad. Sci. USA* **102**, 12321–12325 (2005).
 15. Irimia-Vladu, M., Glowacki, E. D., Voss, G., Bauer, S. & Sariciftci, N. S. Green and biodegradable electronics. *Mater. Today* **15**, 340–346 (2012).
 16. Gamota, D. R., Brazis, P., Kalyanasundaram, K. & Zhang, J. *Printed Organic and Molecular Electronics* (Springer Science & Business Media, 2013).
 17. Benight, S. J., Wang, C., Tok, J. B. H. & Bao, Z. Stretchable and self-healing polymers and devices for electronic skin. *Prog. Polym. Sci.* **38**, 1961–1977 (2013).
 18. Wallace, G. G., Moulton, S. E. & Clark, G. M. Electrode–cellular interface. *Science* **324**, 185–186 (2009).
 19. Martin, D. C. & Malliaras, G. G. Interfacing electronic and ionic charge transport in bioelectronics. *ChemElectroChem* **3**, 686–688 (2016).
 20. Simon, D. T. *et al.* Organic electronics for precise delivery of neurotransmitters to modulate mammalian sensory function. *Nature Mater.* **8**, 742–746 (2009).
 21. Wessling, B. New insight into organic metal polyaniline morphology and structure. *Polymers* **2**, 786–798 (2010).
 22. Groenendaal, L. B., Jonas, F., Freitag, D., Pielartzik, H. & Reynolds, J. R. Poly(3,4-ethylenedioxythiophene) and its derivatives: past, present, and future. *Adv. Mater.* **12**, 481–494 (2000).
 23. Worfolk, B. J. *et al.* Ultrahigh electrical conductivity in solution-sheared polymeric transparent films. *Proc. Natl Acad. Sci. USA* **112**, 14138–14143 (2015).
 24. Lee, K. *et al.* Metallic transport in polyaniline. *Nature* **441**, 65–68 (2006).
 25. Luo, C. *et al.* General strategy for self-assembly of highly oriented nanocrystalline semiconducting polymers with high mobility. *Nano Lett.* **14**, 2764–2771 (2014).
 26. Yuan, Y. *et al.* Ultra-high mobility transparent organic thin film transistors grown by an off-centre spin-coating method. *Nature Commun.* **5**, 3005 (2014).
 27. Sundar, V. C. *et al.* Elastomeric transistor stamps: reversible probing of charge transport in organic crystals. *Science* **303**, 1644–1646 (2004).
 28. Kaltenbrunner, M. *et al.* An ultra-lightweight design for imperceptible plastic electronics. *Nature* **499**, 458–463 (2013).
This paper reports the first ultrathin, integrated, organic transistor circuit and sensor, made on plastic foils with a thickness of 1 µm, in what is now referred to as 'imperceptible electronics'.
 29. White, M. S. *et al.* Ultrathin, highly flexible and stretchable PLEDs. *Nature Photonics* **7**, 811–816 (2013).
 30. Lipomi, D. J., Tee, B. C. K., Vosgueritchian, M. & Bao, Z. Stretchable organic solar cells. *Adv. Mater.* **23**, 1771–1775 (2011).
 31. Lipomi, D. J. *et al.* Skin-like pressure and strain sensors based on transparent elastic films of carbon nanotubes. *Nature Nanotechnol.* **6**, 788–792 (2011).
 32. Matsuhisa, N. *et al.* Printable elastic conductors with a high conductivity for electronic textile applications. *Nature Commun.* **6**, 7461 (2015).
 33. Kim, Y. *et al.* Stretchable nanoparticle conductors with self-organized conductive pathways. *Nature* **500**, 59–63 (2013).
 34. Liang, J. *et al.* Intrinsically stretchable and transparent thin-film transistors based on printable silver nanowires, carbon nanotubes and an elastomeric dielectric. *Nature Commun.* **6**, 7647 (2015).
 35. Park, M. *et al.* Highly stretchable electric circuits from a composite material of silver nanoparticles and elastomeric fibres. *Nature Nanotechnol.* **7**, 803–809 (2012).
 36. Sekitani, T. *et al.* A rubberlike stretchable active matrix using elastic conductors. *Science* **321**, 1468–1472 (2008).
 37. Vosgueritchian, M., Lipomi, D. J. & Bao, Z. Highly conductive and transparent PEDOT:PSS films with a fluorosurfactant for stretchable and flexible transparent electrodes. *Adv. Funct. Mater.* **22**, 421–428 (2012).
 38. Oh, J. Y., Kim, S., Baik, H.-K. & Jeong, U. Conducting polymer dough for deformable electronics. *Adv. Mater.* **28**, 4455–4461 (2016).
 39. O'Connor, B. *et al.* Correlations between mechanical and electrical properties of polythiophenes. *ACS Nano* **4**, 7538–7544 (2010).
 40. Savagatrup, S., Printz, A. D., O'Connor, T. F., Zaretski, A. V. & Lipomi, D. J. Molecularly stretchable electronics. *Chem. Mater.* **26**, 3028–3041 (2014).
 41. Lipomi, D. J. & Bao, Z. Stretchable, elastic materials and devices for solar energy conversion. *Energy Environ. Sci.* **4**, 3314–3328 (2011).
 42. Shin, M. *et al.* Highly stretchable polymer transistors consisting entirely of stretchable device components. *Adv. Mater.* **26**, 3706–3711 (2014).
 43. Chortos, A. *et al.* Mechanically durable and highly stretchable transistors employing carbon nanotube semiconductor and electrodes. *Adv. Mater.* **28**, 4441–4448 (2016).
 45. Bettinger, C. J. & Bao, Z. Organic thin-film transistors fabricated on resorbable biomaterial substrates. *Adv. Mater.* **22**, 651–655 (2010).
 46. Irimia-Vladu, M. "Green" electronics: biodegradable and biocompatible materials and devices for sustainable future. *Chem. Soc. Rev.* **43**, 588–610 (2014).
 47. Campana, A., Cramer, T., Simon, D. T., Berggren, M. & Biscarini, F. Electrocardiographic recording with conformable organic electrochemical transistor fabricated on resorbable bioscaffold. *Adv. Mater.* **26**, 3874–3878 (2014).
 48. Tobjork, D. & Osterbacka, R. Paper electronics. *Adv. Mater.* **23**, 1935–1961 (2011).
 49. Zheng, Y. F., Gu, X. N. & Witte, F. Biodegradable metals. *Mater. Sci. Eng. R.* **77**, 1–34 (2014).
 50. Hwang, S.-W. *et al.* High-performance biodegradable/transient electronics on biodegradable polymers. *Adv. Mater.* **26**, 3905–3911 (2014).
 51. Yu, K. J. *et al.* Bioresorbable silicon electronics for transient spatiotemporal mapping of electrical activity from the cerebral cortex. *Nature Mater.* **15**, 782–791 (2016).
 52. Kang, S.-K. *et al.* Bioresorbable silicon electronic sensors for the brain. *Nature* **530**, 71–76 (2016).
 53. Tao, H. *et al.* Silk-based resorbable electronic devices for remotely controlled therapy and *in vivo* infection abatement. *Proc. Natl Acad. Sci. USA* **111**, 17385–17389 (2014).
 54. Rivers, T. J., Hudson, T. W. & Schmidt, C. E. Synthesis of a novel, biodegradable electrically conducting polymer for biomedical applications. *Adv. Funct. Mater.* **12**, 33–37 (2002).
 55. Yang, Y. & Urban, M. W. Self-healing polymeric materials. *Chem. Soc. Rev.* **42**, 7446–7467 (2013).
 56. Williams, K. A., Boydston, A. J. & Bielawski, C. W. Towards electrically conductive, self-healing materials. *J. R. Soc. Interface* **4**, 359–362 (2007).
 57. Tee, B. C. K., Wang, C., Allen, R. & Bao, Z. An electrically and mechanically self-healing composite with pressure- and flexion-sensitive properties for electronic skin applications. *Nature Nanotechnol.* **7**, 825–832 (2012).
 58. Wang, C. *et al.* Self-healing chemistry enables the stable operation of silicon microparticle anodes for high-energy lithium-ion batteries. *Nature Chem.* **5**, 1042–1048 (2013).
 59. Gong, C. *et al.* A healable, semitransparent silver nanowire–polymer composite conductor. *Adv. Mater.* **25**, 4186–4191 (2013).
 60. Kaltenbrunner, M. *et al.* Ultrathin and lightweight organic solar cells with high flexibility. *Nature Commun.* **3**, 770 (2012).
 61. Myny, K. *et al.* An 8-bit, 40-instructions-per-second organic microprocessor on plastic foil. *IEEE J. Solid-State Circuits* **47**, 284–291 (2012).
 62. Gelincik, G. H. *et al.* Flexible active-matrix displays and shift registers based on solution-processed organic transistors. *Nature Mater.* **3**, 106–110 (2004).
 63. Naber, R. C. G. *et al.* High-performance solution-processed polymer ferroelectric field-effect transistors. *Nature Mater.* **4**, 243–248 (2005).
 64. Subramanian, V. *et al.* Progress toward development of all-printed RFID tags: materials, processes, and devices. *Proc. IEEE* **93**, 1330–1338 (2005).
 65. Khodagholy, D. *et al.* High transconductance organic electrochemical transistors. *Nature Commun.* **4**, 2133 (2013).
 66. Mannsfeld, S. C. B. *et al.* Highly sensitive flexible pressure sensors with microstructured rubber dielectric layers. *Nature Mater.* **9**, 859–864 (2010).
 67. Schwartz, G. *et al.* Flexible polymer transistors with high pressure sensitivity for application in electronic skin and health monitoring. *Nature Commun.* **4**, 1859 (2013).
 68. Yokota, T. *et al.* Ultraflexible, large-area, physiological temperature sensors for multipoint measurements. *Proc. Natl Acad. Sci. USA* **112**, 14533–14538 (2015).
 69. Jeon, J., Lee, H.-B.-R. & Bao, Z. Flexible wireless temperature sensors based on Ni microparticle-filled binary polymer composites. *Adv. Mater.* **25**, 850–855 (2013).
 70. Mulla, M. Y. *et al.* Capacitance-modulated transistor detects odorant binding protein chiral interactions. *Nature Commun.* **6**, 6010 (2015).
 71. Ramuz, M., Hama, A., Rivnay, J., Leleux, P. & Owens, R. M. Monitoring of cell layer coverage and differentiation with the organic electrochemical transistor. *J. Mater. Chem. B* **3**, 5971–5977 (2015).
 72. Lochner, C. M., Khan, Y., Pierre, A. & Arias, A. C. All-organic optoelectronic sensor for pulse oximetry. *Nature Commun.* **5**, 5745 (2014).
This paper reports that pulse oximetry has been achieved in solution-processable organic light-emitting diodes and organic photodetectors, expanding the applications of organic photonic devices to chemical sensing.
 73. Yokota, T. *et al.* Ultraflexible organic photonic skin. *Sci. Adv.* **2**, e1501856 (2016).
 74. Sirringhaus, H. *et al.* High-resolution inkjet printing of all-polymer transistor circuits. *Science* **290**, 2123–2126 (2000).
 75. Fukuda, K. *et al.* Fully-printed high-performance organic thin-film transistors and circuitry on one-micron-thick polymer films. *Nature Commun.* **5**, 4147 (2014).
 76. Reichert, W. M. *Indwelling Neural Implants: Strategies For Contending With The In Vivo Environment* (CRC Press, 2007).
 77. Asplund, M., Nyberg, T. & Inganäs, O. Electroactive polymers for neural interfaces. *Polym. Chem.* **1**, 1374–1391 (2010).
 78. Kozai, T. D. Y. & Kipke, D. R. Insertion shuttle with carboxyl terminated self-assembled monolayer coatings for implanting flexible polymer neural probes in the brain. *J. Neurosci. Meth.* **184**, 199–205 (2009).
 79. Mineev, I. R. *et al.* Electronic dura mater for long-term multimodal neural interfaces. *Science* **347**, 159–163 (2015).
This study expands the stability and clinical benefits of flexible electronics by creating an 'e-dura' that uses stretchable electrodes and microfluidic channels for controllable drug delivery.
 80. Ludwig, K. A., Uram, J. D., Yang, J., Martin, D. C. & Kipke, D. R. Chronic neural recordings using silicon microelectrode arrays electrochemically deposited with a poly(3,4-ethylenedioxythiophene) (PEDOT) film. *J. Neural Eng.* **3**, 59–70 (2006).
 81. Venkatraman, S. *et al.* In vitro and In vivo evaluation of PEDOT microelectrodes

- for neural stimulation and recording. *IEEE Trans. Neural Syst. Rehabil. Eng.* **19**, 307–316 (2011).
81. Green, R. & Abidian, M. R. Conducting polymers for neural prosthetic and neural interface applications. *Adv. Mater.* **27**, 7620–7637 (2015).
 82. Khodagholy, D. *et al.* In vivo recordings of brain activity using organic transistors. *Nature Commun.* **4**, 1575 (2013).
- This paper exploits the transconductance, mechanical flexibility and biocompatibility of organic electrochemical transistors to create a sensor with a high signal-to-noise ratio that records brain activity.**
83. Khodagholy, D. *et al.* NeuroGrid: recording action potentials from the surface of the brain. *Nature Neurosci.* **18**, 310–315 (2015).
 84. Xu, L. *et al.* Materials and fractal designs for 3D multifunctional integumentary membranes with capabilities in cardiac electrotherapy. *Adv. Mater.* **27**, 1731–1737 (2015).
 85. Park, D.-W. *et al.* Graphene-based carbon-layered electrode array technology for neural imaging and optogenetic applications. *Nature Commun.* **5**, 5258 (2014).
 86. Ouyang, L., Shaw, C. L., Kuo, C.-c., Griffin, A. L. & Martin, D. C. In vivo polymerization of poly (3, 4-ethylenedioxythiophene) (PEDOT) in the living rat hippocampus does not cause a significant loss of performance in a delayed alternation (DA) task. *J. Neural Eng.* **11**, 026005 (2014).
 87. Hassarati, R. T., Marcal, H., Foster, L. J. R. & Green, R. A. Biofunctionalization of conductive hydrogel coatings to support olfactory ensheathing cells at implantable electrode interfaces. *J. Biomed. Mater. Res. B* **104**, 712–722 (2016).
 88. Richardson, R. T. *et al.* Polypyrrole-coated electrodes for the delivery of charge and neurotrophins to cochlear neurons. *Biomaterials* **30**, 2614–2624 (2009).
 89. Jonsson, A. *et al.* Therapy using implanted organic bioelectronics. *Sci. Adv.* **1**, e1500039 (2015).
 90. Williamson, A. *et al.* Controlling epileptiform activity with organic electronic ion pumps. *Adv. Mater.* **27**, 3138–3144 (2015).
 91. Schmidt, C. E., Shastri, V. R., Vacanti, J. P. & Langer, R. Stimulation of neurite outgrowth using an electrically conducting polymer. *Proc. Natl Acad. Sci. USA* **94**, 8948–8953 (1997).
 92. Huang, J. *et al.* Electrical stimulation to conductive scaffold promotes axonal regeneration and remyelination in a rat model of large nerve defect. *PLoS ONE* **7**, e39526 (2012).
 93. Wong, J. Y., Langer, R. & Ingber, D. E. Electrically conducting polymers can noninvasively control the shape and growth of mammalian-cells. *Proc. Natl Acad. Sci. USA* **91**, 3201–3204 (1994).
 94. Wan, A. M.-D. *et al.* 3D conducting polymer platforms for electrical control of protein conformation and cellular functions. *J. Mater. Chem. B* **3**, 5040–5048 (2015).
 95. Ghezzi, D. *et al.* A polymer optoelectronic interface restores light sensitivity in blind rat retinas. *Nature Photonics* **7**, 400–406 (2013).
 96. Qiu, F. *et al.* Magnetic helical microswimmers functionalized with lipoplexes for targeted gene delivery. *Adv. Funct. Mater.* **25**, 1666–1671 (2015).
 97. Tee, B. C. K. *et al.* A skin-inspired organic digital mechanoreceptor. *Science* **350**, 313–316 (2015).
- By mimicking the characteristics of animal skin and nerve cells, this study integrates ultrasensitive pressure sensors with organic transistor circuits to stimulate mouse brain.**
98. Xu, S. *et al.* Soft microfluidic assemblies of sensors, circuits, and radios for the skin. *Science* **344**, 70–74 (2014).
 99. Takei, K. *et al.* Nanowire active-matrix circuitry for low-voltage macroscale artificial skin. *Nature Mater.* **9**, 821–826 (2010).
 100. Someya, T. Building bionic skin. *IEEE Spectrum* **50**, 50–56 (2013).
 101. Irimia-Vladu, M. *et al.* Indigo – a natural pigment for high performance ambipolar organic field effect transistors and circuits. *Adv. Mater.* **24**, 375–380 (2012).
 102. Buchko, C. J., Slattery, M. J., Kozloff, K. M. & Martin, D. C. Mechanical properties of biocompatible protein polymer thin films. *J. Mater. Res.* **15**, 231–242 (2000).
 103. Shackelford, J. F., Han, Y.-H., Kim, S. & Kwon, S.-H. *CRC Materials Science and Engineering Handbook* (CRC Press, 2016).
 104. Kim, D.-H. *et al.* Epidermal electronics. *Science* **333**, 838–843 (2011).
- This study integrates inorganic electronic elements into an ultrathin, low-modulus device that conforms to the surface of skin, leading to monitoring of vital information without the need for adhesives.**
105. Webb, R. C. *et al.* Ultrathin conformal devices for precise and continuous thermal characterization of human skin. *Nature Mater.* **12**, 938–944 (2013).
 106. Park, S. I. *et al.* Soft, stretchable, fully implantable miniaturized optoelectronic systems for wireless optogenetics. *Nature Biotechnol.* **33**, 1280–1286 (2015).
 107. Viventi, J. *et al.* Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. *Nature Neurosci.* **14**, 1599–1605 (2011).

Acknowledgements The authors acknowledge R. Nawrocki for fruitful discussions and J. Xu for help with formatting.

Author information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/2gxxtnj. Correspondence should be addressed to T.S. (someya@ee.t.u-tokyo.ac.jp).

Mimicking biological functionality with polymers for biomedical applications

Jordan J. Green & Jennifer H. Elisseeff

The vast opportunities for biomaterials design and functionality enabled by mimicking nature continue to stretch the limits of imagination. As both biological understanding and engineering capabilities develop, more sophisticated biomedical materials can be synthesized that have multifaceted chemical, biological and physical characteristics designed to achieve specific therapeutic goals. Mimicry is being used in the design of polymers for biomedical applications that are required locally in tissues, systemically throughout the body, and at the interface with tissues.

The footpads of a gecko. The adhesive used by mussels to cling to rocks. The infectivity of a virus. These are only a few of the structures and processes that scientists and engineers are striving to mimic as they design materials to solve a host of problems. The natural world has inspired us for centuries, and today's understanding of biological structures and the functions of complex physiological systems is providing a wealth of models to help materials scientists and engineers build new therapeutics.

An improved molecular-level understanding of microenvironments, biological nanoparticulates and multivalent macromolecular assemblies is leading to the biologically inspired creation of materials as diverse as synthetic hydrogels that recapitulate the extracellular matrix, morphologically matched, polymeric, artificial pathogens, and biomimetic adhesives that bind tightly in wet environments. A unifying theme in these designs is that they emulate both the physical properties of core materials and the chemical properties of biological surfaces to create biocompatible, artificial, polymer-based interfaces that can direct the functions of biological cells in increasingly complex ways.

In this Review we discuss how these biomimetic polymeric systems can be used for a diverse range of therapeutic biomedical applications in tissues, systemically throughout the body, and at the interface with tissues. Polymeric materials can be designed to mimic local tissue properties. The chemical composition and postsynthesis materials processing define the structural and mechanical properties, while sequestered proteins and bioactive surfaces produce the functionality of implants. Polymeric materials for systemic function include synthetic viruses for gene delivery, synthetic particles with high aspect ratios for immunotherapy, and structures that mimic a range of cells from platelets to dendritic cells. At the interface with natural tissues, multivalent polymers can mimic both adhesive and highly lubricated surfaces found in nature. Thorough evaluation of the way these artificial biomaterials interact with cells and tissues will improve their design and lead to therapeutic polymeric materials that have the potential to benefit the lives of millions of people.

Localized tissue function

Hydrogels are a class of biomaterials that serve as a useful tool and building block for creating structures that mimic nature. Indeed, the tissues found in plants and animals are themselves composed largely of hydrogels. They are formed by crosslinking polymers — macromolecules composed of smaller repeating units called monomers — to create an insoluble network that can absorb water without dissolving. In natural

biological systems, proteins and proteoglycans are the polymers that make up the building blocks of hydrogels, which in turn form tissue structures. Synthetic polymers can be used to form hydrogels that mimic biological systems, and biological signals, such as growth factors and cell-surface receptors, can be incorporated in the hydrogel to modulate the response of cells and tissues.

Hydrogels as tissue mimics

Hydrogels mimic many of the properties of tissues in the body, so a straightforward biomedical application has been to use them as vessels for culturing cells, and also as scaffolds for tissue engineering to rebuild and repair tissues *in vivo*. The physical and biological requirements of hydrogels for these applications are many and varied, and they can be modified in many ways to mimic the natural tissue environment. By controlling the polymer chemistry and crosslinking density, the physical properties of a hydrogel, such as its water content, mechanical strength and elasticity, can be manipulated to resemble those of particular natural tissues or to trigger a desired biological outcome¹ (Fig. 1a,b). For instance, culturing stem cells on harder or softer hydrogel surfaces encourages them to become more like bone or brain, respectively².

Hydrogels can also be formed in specific geometries to create a particular biomimetic structure or to control cell shape and function. Restricting mammary epithelial cells to confined spaces of different shapes, for example, varies their ability to form complex developmental structures, such as breast acini³, the smallest lobule structures in the gland. Biological cues such as active peptides, proteins and proteoglycans can be integrated into hydrogels through both covalent linkages and non-covalent interactions to create a synthetic mimic of the natural extracellular matrix (ECM), which is the structural and biological framework for all tissues in the body. The composition of the hydrogel itself can even be derived from natural sources, as in silk-based materials⁴. With appropriate choice of the polymeric building blocks and the use of crosslinking chemistry, hydrogels can be highly biocompatible, and so can be used as two-dimensional vehicles for cell culture, or as three-dimensional microenvironments, with cells being seeded on top or encapsulated during the crosslinking process when the hydrogel is formed.

Dynamic hydrogels

An important aspect of the native tissue environment is the 'dynamic reciprocity' of the cell with its ECM. We are learning more about the complexity of the signals that cells can sense in their local

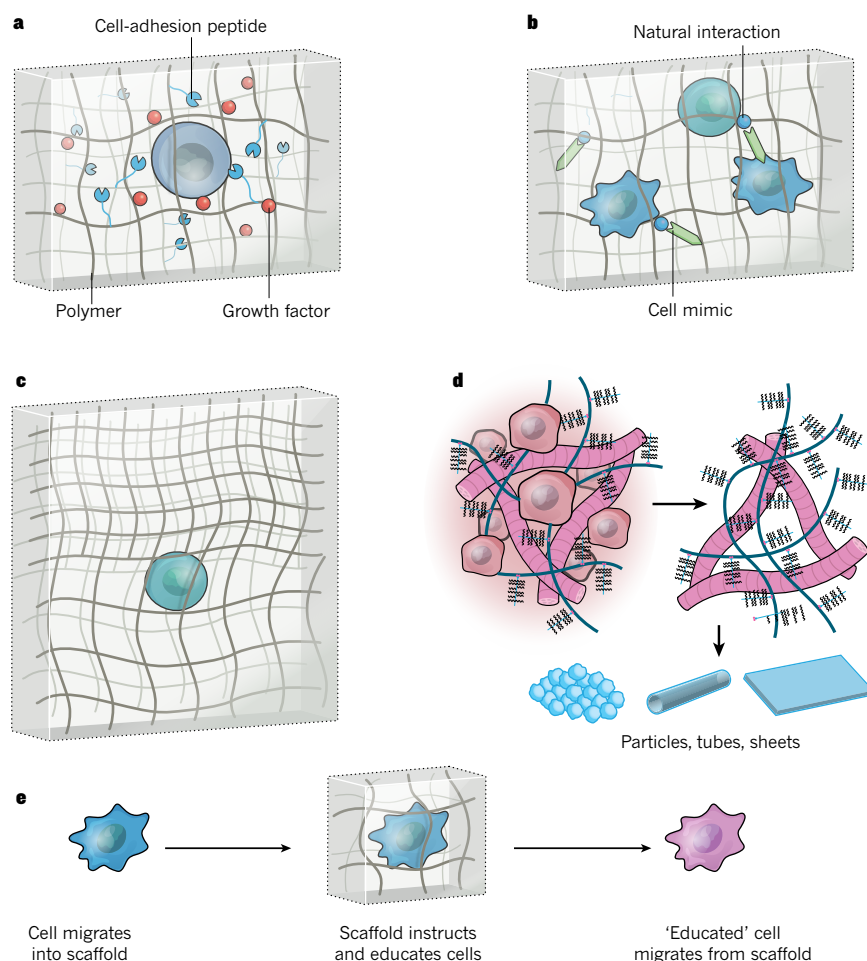


Figure 1 | Strategies to create synthetic environments that mimic tissues. **a**, Cell-matrix mimics. Synthetic hydrogels made of polymers can be modified with peptides or proteins (such as growth factors) and cell-sensitive degradable crosslinks that mimic many of the properties of the native tissue extracellular matrix (ECM)¹. **b**, Cells often live in communities, so it can be useful to mimic cells by attaching surface proteins to a hydrogel⁶. **c**, The mechanical properties of the hydrogel can be controlled by varying the crosslinking density or using chemistries that change the mechanical properties independently of ligand presentation; by metal ligand chemistry from the mussel⁶⁸; or by varying the ionic crosslinking density in alginate hydrogels derived from seaweed²². **d**, The natural ECM from tissues can be processed to remove cells, and the remaining matrix can be processed into different scaffold forms, such as particles, tubes and sheets. **e**, A cell can migrate into a polymeric material where it is 'educated' by signals embedded in the scaffold before leaving to perform a particular function in the body. (Figure reproduced with permission from refs 1, 6, 22 and 68.)

microenvironment, and these can in turn be built into scaffolds. The simplest and earliest strategy for incorporating biology into a hydrogel was the inclusion of pieces of the natural environment, which could be large intact molecules, or only the active region of a molecule, or even more fundamental small molecules and chemical functionalities, all of which can affect cell responses such as adhesion, migration and phenotype. For example, chemical modification to incorporate the functional groups *t*-butyl, phosphate or carboxylic acid into a purely synthetic poly(ethylene glycol) (PEG) hydrogel has been shown to stimulate the differentiation of human mesenchymal stem cells (hMSCs) towards adipogenic, osteogenic and chondrogenic pathways, respectively⁵.

Hydrogels are frequently considered to be mimics of the ECM, but they can also include signals to replicate interactions between cells. For example, the EphA–EphrinA interaction between islet cells has been replicated in microcapsules with low cell density. The low cell density ensures better nutrient availability for the cells, and the artificial cell signal 'tricks' the cells into sensing the high cell density they prefer, thereby enhancing their survival⁶ (Fig. 1b). Biopolymers in a hydrogel can be manipulated and degraded by cells as they would be in the native ECM, but specific cell-controlled degradation can also be built into synthetic hydrogels through enzyme-sensitive crosslinks. These crosslinks have been tailored to promote everything from controlled cell movement to slowed resorption of materials. For instance, specific peptide crosslinkers incorporated in hydrogels increased bone formation in cranial defects in mice⁷. Controlled degradation can come into play when attempting to prolong a healing effect, for example in wound dressings, or when slowing the delivery profile of an encapsulated drug or cell payload. This extensive toolbox of materials with specific chemical, biological and physical cues can be embedded into biomaterials, so interdisciplinary teams are

now needed to define the important biological areas to be tackled, and to determine the clinical need and any constraints that must be addressed.

Biological cues

Although the interaction of cells directly with their environment (synthetic or natural) and with each other is an important functional and therapeutic target, the ECM also has an often overlooked function: the binding, storage and control of growth-factor signalling activity (Fig. 1a). Vascular endothelial growth factor (VEGF) has been physically entrapped during the polymerization of PEG hydrogels containing enzymatically degradable peptide sequences, and released in response to cell-secreted proteases and increased vascular development in both *in vitro* cell assays and in mice⁸. Growth factors can also be modified to bind covalently and non-covalently to synthetic and native ECM scaffolds. For example, growth factors modified with the ECM-binding region of placental growth factor were able to bind multiple ECM proteins, including vitronectin, tenascin C and type I collagen, that are found in tissues or made available in synthetic materials⁹. VEGF that has been modified with these ECM-binding motifs improved healing in chronic diabetic wounds and bone defects. Covalent integration of full-length proteins (including vitronectin) into PEG hydrogels can be accomplished with temporal and spatial specificity by various photoreversible means. PEG hydrogels have been synthesized with photo-responsive, hetero-bifunctional linkers, yielding an aldehyde-reactive alkoxyamine when exposed to light. Reactive aldehyde moieties substituted on target proteins are then able to bond covalently to the hydrogel matrix when and where light exposure occurs, and can control cell function. For instance, the differentiation of hMSCs into bone tissue was limited to regions where vitronectin was immobilized with photosensitive linkers in the hydrogel. When

the protein was cleaved from the hydrogel, cells subsequently scaled back their expression of proteins that are specific for bone growth¹⁰.

The mechanical environment of the ECM can also be fine-tuned in a hydrogel. The biophysical properties of the base polymer that makes up the hydrogel, and the number and type of crosslinks between the polymer chains, can all affect the hydrogel's mechanical properties (Fig. 1c). These properties are now known to have a large effect on cell behaviour. The material's elasticity can affect stem-cell differentiation, cell shape, the efficiency of gene transfection, the immune response, and many other biological properties. It has recently been found that cells react differently to elastic and viscoelastic hydrogels (which have properties of both a liquid and a solid), as this changes the relaxation behaviour of the material after cells interact, altering the cells' response¹¹. Cells can also sense and respond to heterogeneous variations in the mechanical properties of hydrogels. Bundles of type I collagen combined with polyacrylamide during free-radical polymerization produced hydrogels with regions of enhanced stiffness that mimicked scar tissue and induced MSCs to express smooth-muscle actin, further enhancing fibrotic responses¹² and collagen deposition.

Biological cues in the form of peptides can be presented to cells alongside various independently controlled mechanical properties to determine their combined roles in disease processes. In cancer, for example, mechanics alone does not control malignant-cell migration in alginate or PEG hydrogels. But a combination of integrating cell adhesion through peptides bound to the hydrogels, and manipulating the mechanics by changing the crosslinking density, had an effect^{13,14}. Biologists are now uncovering the mechanical sensing properties of cells. For example, the Hippo pathway, which coordinates several cell processes through cytoskeletal tension, was discovered to be associated with embryonic patterning and organ size, and is now linked to cell proliferation and wound healing¹⁵. The Hippo–Yap pathway has been manipulated by using soft poly(dimethylsiloxane) (PDMS) materials to enhance the neuronal differentiation of stem cells¹⁶. Biologists and engineers can now come together to design hydrogels that modulate these developmental signals to further control cell behaviour for therapeutic benefit.

External hydrogel manipulation

The mechanical and biological manipulation of hydrogels can be controlled spatially, so specific signals can be displayed in a pattern that mimics the morphological gradients found in development and wound healing. External tools such as light or temperature further expand hydrogel manipulation to the temporal scale, so changes can be made to the structure and presentation of biophysical signals on demand. By using light, biological signals can be attached to or released from a hydrogel, and structural crosslinks can be made or broken to respectively stiffen or weaken a material in real time. Light that has penetrated skin tissue and reached an underlying hydrogel has been used to expose cell-adhesive peptides, which was found to raise the number of inflammatory cells and increase fibrosis in a PEG implant. When these cellular changes were combined with protease-sensitive degradation of the hydrogel and the release of vascular endothelial growth factor (VEGF), vascularization of the implant increased⁸. Photocleavable cell-adhesion peptides can also provide a dynamic cell signal on biomimetic peptide amphiphilic materials. These scaffolds can be processed into nanofibres or hydrogels, depending on the desired application¹⁷. In PEG hydrogels, the presence of the adhesion peptide RGD (arginine, glycine, aspartic acid) initially enhanced the generation of cartilage from MSCs, but later inhibited it. Light-based cleavage of RGD from hydrogels after one week of culture was able to increase it again, highlighting the need for dynamic biological signals^{15,18}. And multiple signals can be controlled independently with photosensitive bonds that respond to different wavelengths of light.

These tunable, biomimetic environments can tell us a great deal about cell behaviour in specific microenvironments *in vitro*, but the connection to therapeutic responses in animals (or people) is less clear, and few technologies that use these dynamic scaffolds have reached clinical

testing. Furthermore, hydrogels that comprise numerous biological elements and chemical substitutions designed to mimic native tissues can be difficult to manufacture and may have complex regulatory pathways. However, information gleaned from *in vitro* studies of structure and function can still aid the design of the biomimetic structures needed to modulate local cell behaviour for applications such as wound healing in simplified systems that are amenable to clinical translation.

An artificial ECM can be built from the bottom up using synthetic polymers as described above, whereas biological scaffolds derived from the ECM are engineered from the top down using tissues. Biological scaffolds, composed of processed native ECM tissue, have already been implanted in millions of people for applications ranging from hernia treatment and repair of the rotator cuff in the shoulder to breast reconstruction during surgery. Biological scaffolds are created by killing cells and washing out their remnants, leaving the tissue ECM, which is a complex mixture of proteoglycans and proteins that can be processed into implants, injectable hydrogels or particle suspensions, depending on the clinical need, where they serve as a scaffold for cell migration and new tissue development¹⁹ (Fig. 1d). Each specific tissue is processed in unique conditions to optimize the removal of cells while minimizing changes and damage to the composition and organization of the ECM. Although these materials do not have the highly controlled composition of synthetic scaffolds, they have shown remarkable clinical efficacy. A wide range of biological scaffolds derived from different tissue sources are in preclinical, clinical and commercial use today²⁰ (Table 1). Tissue ECMs closely resemble the native tissue from which they are derived and are also likely to provide important, as yet unknown, biological signals that attract and modulate cell function. This dual structure–function relationship can be harnessed to use these ECMs for regenerative medicine. For instance, introducing porcine ECM to a wound not only fills a tissue void but also activates pathways that augment the wound-healing response by attracting the cells responsible for tissue reconstruction into the material and subsequently altering their behaviour^{20,21} (Fig. 1e). Implantation and the regenerative response to the biological scaffolds, regardless of their tissue source, have been associated with specific immune signatures that are now understood to be an early critical step in the repair and regeneration process²¹. So defining and actively triggering the physiological processes (such as the immune response) that are driving a desired outcome (such as regeneration) may be more important than mimicking the precise tissue structure in a biomaterial. This finding should inform future efforts to design therapeutic hydrogels.

Systemic function

Just as these macroscale polymeric hydrogels are being designed to mimic hydrated biological microenvironments before being administered locally, complementary polymeric nanoscale and microscale constructs are also being designed, but these are used to mimic biological activity through systemic administration. Natural immunological agents are increasingly being reproduced through the use of synthetic bio-inspired polymeric materials (Fig. 2). The growing success of cancer immunotherapy in particular has spurred interest in using biomaterials to mimic elements, or modulate components, of the immune system. On the macroscale, combinations of cytokines with biomaterial scaffolds²² and injectable materials²³ can be used to stimulate dendritic (antigen-presenting) cells, enhancing the efficiency of cancer vaccines — a platform that is now being tested in clinical trials²⁴. They work in a similar way to regenerative scaffolds, as cells migrate to the immunomodulatory biomaterial where they are 'educated' to perform a new function (Fig. 1e). Taking inspiration from the immune stimulation that occurs when tissues are wounded, alternatively spliced fibronectin domains that serve as Toll-like-receptor agonists have been incorporated into hydrogels to increase the responses of CD8-positive T cells in mouse tumour models²⁵. In terms of designing materials to enhance biological compatibility, the size of implanted spheres has been shown to dramatically affect fibrosis and 'foreign body' reactions in a range

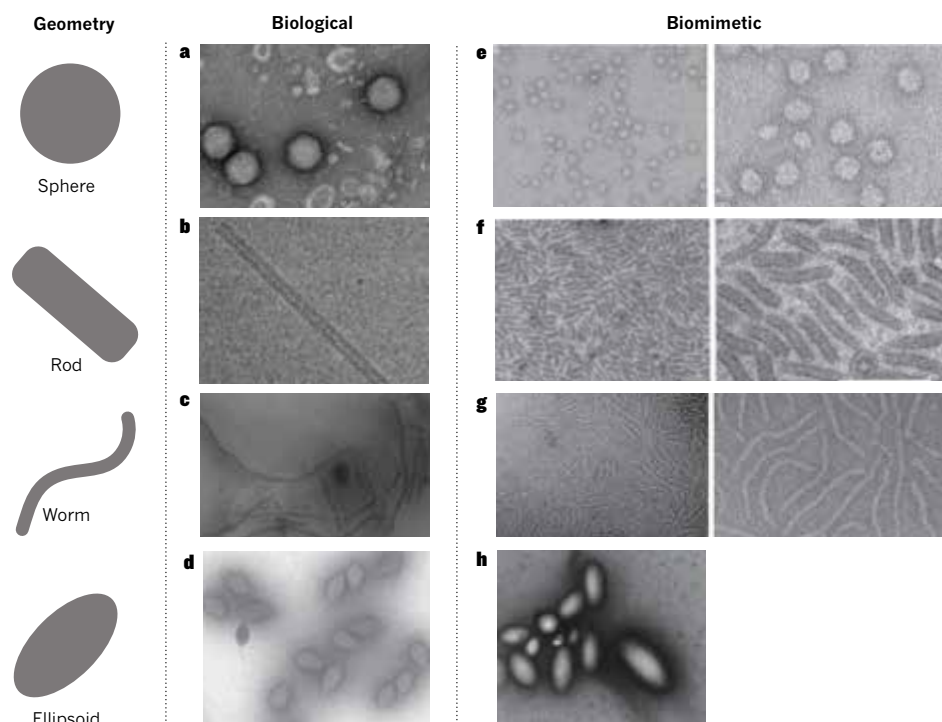


Figure 2 | Biomimetic polymeric nanostructures can be constructed to mimic the geometries of biological viruses for systemic delivery. Viruses can have spherical, rod-like, worm-like and ellipsoidal shapes, as shown here: **a**, adenovirus³³; **b**, tobacco mosaic virus³⁴; **c**, Ebola virus³⁶; **d**, *Acidianus convivator*³⁷. **e–h**, These viruses are mimicked respectively by polyethylene glycol-b-polyphosphoramidate/DNA polyelectrolyte complex nanoparticles³⁸ with spherical (**e**), rod-like (**f**) and worm-like (**g**) shapes, and by ellipsoidal poly(lactic-co-glycolic acid)-based nano artificial antigen-presenting cells⁵³ with an aspect ratio of 2 (**h**). Non-spherical shapes are shown to have enhanced efficacy for the intracellular delivery of DNA and the extracellular presentation of protein. (Figures are reproduced with permission from refs 33, 34, 36–38 and 53.)

of material types, including alginate hydrogels, stainless steel, plastics and glass; a length scale of 1.5 mm and greater is most effective^{26,27}. When combined with surface engineering to create a ‘stealthy’ chemical surface^{28,29}, which inhibits both the deposition of collagen and macrophage recognition, there is reduced inflammation and fibrosis. This has enabled encapsulated insulin-producing human stem-cell-derived beta cells to survive for approximately six months *in vivo*, maintaining control of glucose in mice^{30,31}.

On the micro- and nanoscales, synthetic bio-inspired polymeric materials can be designed to mimic natural immunological agents, including pathogens and even cells, in both form and function. But form does not always have to be a literal interpretation of nature to have the desired function. Aircraft are a good example, as they mimic the structure of birds, but their wings are fixed and do not flap. When designing immunomodulatory particles, the two crucial aspects that need to be carefully considered and engineered are the design of both the particle’s core and its surface. As with macroscale hydrogels, the mimicry of its physical properties needs to be combined with mimicry of the chemical properties if the particle is to drive the behaviour of target cells.

Mimicking pathogens

Pathogens, in particular viruses, are known to have diverse geometry and surface properties (Fig. 2). For example, the size and shape of virus particles can vary from approximate spheres of diameter 20 nm (adeno-associated virus)³² to 100 nm (adenovirus; Fig. 2a)³³, to short nanorods 300 nm long (tobacco mosaic virus; Fig. 2b)³⁴, longer nanorods of 700 nm (*Stygiolobus*)³⁵, and still longer worm-like shapes approximately 1 µm long (Ebola virus; Fig. 2c)³⁶. Other anisotropic viral shapes have also been investigated, such as the ellipsoid-shaped *Acidianus convivator*³⁷. The virus particle’s surface may be non-enveloped and charged (for example, adenoviruses) or lipid-enveloped (retroviruses).

A similar diversity of physical features is being engineered in bio-inspired non-viral polymeric systems for gene delivery. For example, DNA-containing polymeric nanoparticles have been made with differential geometry to mimic the morphology of natural viruses. The nanoscale constructs are composed of a cationic co-polymer and anionic DNA that, through bottom-up self-assembly in solvent with varying polarity, can form nanocomplexes³⁸. Spherical nanoparticles (Fig. 2e), nanorods (Fig. 2f) and worm-like shapes (Fig. 2g) have each

been constructed with this bottom-up assembly technique, and emulate the biological structure of viruses as well as their function, which is the intracellular delivery of genes. *In vivo* studies in rats have shown that systemically delivered worm-like particles have enabled gene delivery to the liver that is two and three orders of magnitude higher than exogenous expression from rod-like nanoparticles and nanospheres, respectively³⁸. The biomimicry of physical properties such as geometry enables polymeric nanostructures to reach specific tissues, microenvironments and cells where they can potentially program target cells, including immune cells and stem cells, to follow instructions encoded by exogenous nucleic acids.

Polymeric synthetic pathogens have the potential to drive effective immunization against target pathogens while being safe and non-infectious. One strategy for designing particles is to use lipid-coated polymeric particles with surfaces that display antigen³⁹. This approach enables individual tuning of the core-shell structure, and has largely been investigated using spherical particles. To mimic pathogens for antigen presentation, better ways of tuning the physical properties of the polymeric particle are needed. One approach for immunomodulation is to present antigens on surfaces with different geometries⁴⁰. Using particle replication in a non-wetting template (PRINT) process, top-down mould fabrication results in particles with defined and homogeneous size and shape⁴¹. The same chemical signal (phosphatidylserine) and weight percentage were incorporated into two differently shaped particles — nanorods (80 nm × 320 nm) and cylinders (1 µm × 1 µm × 1 µm) — so that the effect of the geometry could be evaluated. Unlike the cylinders, the nanorods exhibited an anti-inflammatory response when administered to dendritic cells, reducing the subsequent activation of CD4-positive T cells, reducing the production of interferon-γ, increasing the abundance of regulatory T cells, and improving outcomes in an experimental autoimmune encephalitis mouse model for multiple sclerosis⁴⁰. Such bioengineered materials are useful for promoting tolerance in transplantation applications. Through biomimetic design, future particulate systems have the potential to be applied systemically but have a localized and specific function. Localized immunosuppression in a target microenvironment could downregulate the immune response at a transplant site without compromising the function of the immune system overall, unlike conventional immunosuppressants.

The surfaces of biological pathogens and cells are often composed of

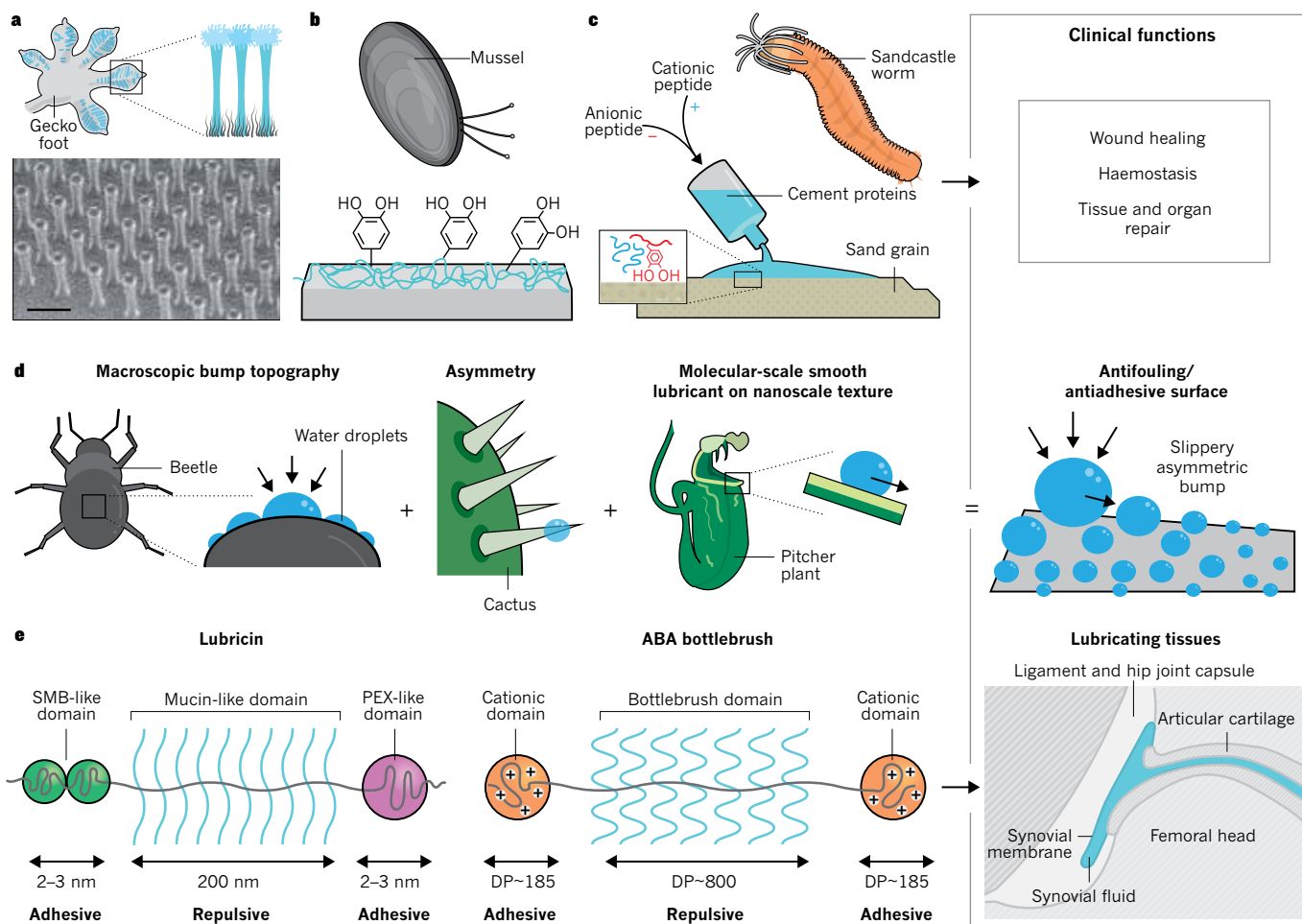


Figure 3 | Biomimetic materials for the design of tissue adhesives and device coatings. **a**, Geckos can walk up walls and hang upside down with the help of pillars and setae on their feet^{67,69}. **b**, Mussels use chemistry that functions in a wet environment to create a powerful adhesive. Mussel coatings can be applied to gecko pillars to create sticky materials that function in a physiological environment^{24,60,61,68}. **c**, Sandcastle worms create 3D structures

in sea water by using adhesives organized into vesicles that can be replicated with biocompatible chemistries to synthesize medical adhesives⁶³. **d**, The surface geometries of beetles, cactus needles and pitcher plants create a hydrophobic layer that can be captured in device coatings^{27,29}. **e**, Lubricin protein can be mimicked by bottlebrush polymers to lubricate tissues such as articular cartilage⁷⁶. DP, degree of polymerization.

lipid membranes, so one intriguing approach is to use naturally derived biological membranes to coat synthetic polymeric surfaces. Membranes from various cell types have been used to coat particles to mimic the surfaces of erythrocytes⁴², cancer cells⁴³ and platelets⁴⁴. Encapsulating polymeric particles in erythrocyte-derived vesicles has enabled them to circulate in mouse blood, extending their half-life compared with fully synthetic methods such as PEGylation⁴². Because these nanostructures mimicked the structure of the erythrocyte surface, they absorbed pore-forming toxins from the blood, extending survival in a mouse model of *Staphylococcus aureus* α -toxin infection⁴⁵. Cancer-cell membranes can similarly be coated on polymeric nanoparticles to enable core-shell nanostructures that mimic natural pathogens and present antigen and deliver adjuvant to dendritic cells — the key cells for directing a systemic antitumour immune response⁴³. Similar top-down approaches have also been used to coat various synthetic particulate surfaces to mimic the biological cell surfaces of leukocytes⁴⁶, macrophages⁴⁷ and mesenchymal stem cells⁴⁸. These approaches allow the precise camouflaging of engineered polymeric particles with the defined chemical and biological surfaces needed for systemic distribution, targeting and function in living organisms.

Mimicking cells

New cellular therapies must overcome major regulatory, manufacturing and cost hurdles if they are to have a significant effect in the clinic, so increased attention is being given to acellular materials that have

cell-like functions. Polymeric micro- and nanoparticulate systems, used as off-the-shelf acellular biological therapies, have the potential for ease, robustness and stability in manufacture and scale-up that is not achievable with cellular systems. Through improved design, they can mimic the function of therapeutic cellular systems, achieving similar biological efficacy with a biodegradable and safe material. Moreover, such polymer-based systems have the ability to purposely not mimic particular aspects of the cells that they emulate if this can improve therapeutic outcomes. For example, biomimetic acellular particles can be designed to ignore inhibitory signals found in the microenvironment that are known to reduce the potency of biological cells; they do not require a supply of oxygen and nutrients to remain potent; they will not differentiate to a different phenotype; and they will not be destroyed by apoptosis shortly after being administered.

Coating polymeric materials with components from the surfaces of naturally occurring particulates such as erythrocytes allows them to be viewed as ‘self’, rather than ‘foreign’. Instead of transferring a full biological membrane onto a polymeric nanostructure, camouflage can be achieved by mimicking a few key signals of self, such as peptides^{49,50}. Long systemic circulation times for nanoparticles have been achieved by conjugating the CD47 mimetic peptide, which acts as the CD47 biological ‘don’t eat me’ signal, to reduce uptake by macrophages⁴⁹. This approach enabled enhanced targeting and labelling of tumours after the systemic injection of polymeric nanoparticles. Intriguingly,

physical properties, including rigidity and shape, can further modulate macrophage uptake, in some cases prevailing over the chemical surface signals for self-altogether⁵¹. This interplay between mimicking the physical and mechanical core properties, and the chemical and topological surface properties, is important for mimicking biological functionality with polymers. Efforts to control this interplay and improve biomimetic design are needed for next-generation therapeutics, because meta-analysis of traditional nanomedicine approaches has recently revealed a lack of efficiency; for example, less than 1% of an administered cancer treatment was found to reach solid tumours⁵².

To mimic the presentation of biological molecules from the surface of biological cells to the extracellular environment, polymeric particles can be constructed on the micrometre scale with matching geometry and surfaces. One cell type that it is important to mimic is the antigen-presenting cell (APC), which directs the immune system. In many ways, APCs, such as dendritic cells, are the conductors of the immune system, sampling both extracellular and intracellular pathogen signals and presenting them to many effector cells along with biological regulatory signals. Artificial APCs can be constructed from biodegradable polymers with surfaces that mimic natural APCs in size, non-spherical shape (Fig. 2h) and protein composition, including a signal 1 that emulates major histocompatibility complex molecules that bind with peptide antigen and T-cell receptors, and a signal 2 that binds co-stimulatory molecules on the T cell^{53–55}, such as CD28. Artificial APCs (aAPCs) have been shown to amplify CD8-positive T cells *in vitro* and *in vivo* in directing the immune system against specific cancer antigens in a mouse model of melanoma⁵⁵. Reproducing the size and the geometry is critical, as microparticle-based aAPCs have superior efficacy to nanoparticle-based aAPCs, and ellipsoidal aAPCs (Fig. 2h), which have a high aspect ratio and a stretched shape, have superior efficacy to spherical aAPCs with the same volume and surface protein content^{53,55}.

Similar biomimetic ‘artificial cell’ approaches can be constructed for non-immune cell functions as well, such as the clotting properties of platelets. Polymeric particulates coated with the natural membranes from platelets are able to bind to damaged vasculature and platelet-adhesive pathogens *in vivo* in rats⁴⁴. Purely synthetic, bottom-up polymer approaches have also been successfully shown to mimic the activity of platelets, if not their structure. An example is a peptide block copolymer that forms by nanoprecipitation into nanoparticles that control bleeding and improve survival when administered intravenously in a rodent model of blast trauma⁵⁶. Platelet-like nanocapsules have also been fabricated using a layer-by-layer technique to construct hollow, flexible, disc-shaped particles with ligand-mimetic surfaces that reduce bleeding time in a tail-amputation mouse model⁵⁷. On a macro scale, self-assembling peptides in hydrogels can form nanofibres in the presence of blood that are similar to the fibres formed in the natural clotting cascade⁵⁸. And incorporating a biological coagulant from snake venom into nanofibres can also augment clot formation, turning a dangerous toxin into a potentially useful therapeutic⁵⁹. Across several fields, mimicking biological functionality with polymers can enable the development of innovative therapeutics for systemic use.

Interfacial function

As well as basing hydrogels and other polymeric structures on human tissues, researchers have hijacked structures in animals ranging from marine worms and molluscs to geckos and beetles as they seek to design materials that can accomplish remarkable feats to serve unmet biomedical needs. An elegant example of this mimicry is found in biomedical adhesives and coatings (Fig. 3). There is a clinical need for tissue adhesives in most areas of medicine and surgery, but they have challenging design requirements. They need efficient tissue adhesion, must be biocompatible, and require mechanical properties that mimic the native tissue, but medical adhesives should also have the ability to operate in the wet *in vivo* environment, and ideally be able to stimulate tissue repair. Marine animals such as molluscs have adhesives that can attach them to almost any substrate in a wet environment, and researchers

have borrowed their chemistry to create adhesive hydrogel materials and coatings (Fig. 3b). Elements of the mussel’s adhesive proteins and their reversible metal coordination and catechol chemistries linked to synthetic polymers have adhesive, self-healing properties in a wet environment^{24,60,61}. These materials have been applied to some of the most challenging wet environments, including *in utero* fetal surgery in pigs, where they were able to bind tissues bathed in amniotic fluid⁶². Sandcastle worms are other sea creatures that produce adhesives in a wet environment (Fig. 3c). These animals secrete vesicles that are rich in protein adhesives that activate when they reach sea water by using oppositely charged moieties to bind surrounding surfaces in the environment to build their house. Using a similar mechanism, complexed polyelectrolytes (catechol-containing anionic polymers and polycations) can bind glass, metal, plastic and biological molecules in a wet environment⁶³. Synthetic particles made of alginate and a hydrophobic light-activating adhesive were designed to mimic the worm vesicles, which aggregate to form a glue, but with the added functionality of being reactive to light⁶⁴. This ‘nanoglue’ has been used to bind both epicardial tissue and detached retinas in *ex vivo* models. The clinical relevance of the glue, and the development of a device to deliver the nanoglue in a beating heart, was validated in a preclinical swine model⁶⁵. As always, efficacy combined with simplicity is the key for successful clinical implementation on a large scale.

Biomedical coatings for medical devices and implants also have challenging *in vivo* requirements owing to their longevity and the increasing demands on performance. The design specifications depend on the clinical application, but in general, materials should be biocompatible and prevent the adhesion of microbes⁶⁶. More sophisticated coatings might promote healing, and perhaps be self-healing in the case of damage or wear. Devices and implants are moving from minimal biological integration and surface design to having the ability for specific molecular and biological interactions that will eventually result in new functions and enhanced performance.

The gecko is a remarkable animal that has inspired the design of adhesives and coatings owing to its ability to walk on vertical surfaces and even stick to ceilings⁶⁷ (Fig. 3a). Gecko feet have a dense array of projections that are covered in nanoscale setae, which use physical interactions such as van der Waals forces and capillary action to adhere to surfaces. The shape of the gecko setae, combined with coating chemistry based on the mussel adhesive, has produced surfaces that can reversibly adhere in both wet and dry environments⁶⁸. The

Table 1 | Biomimetic biological scaffolds engineered from tissues²⁰

Development stage	Tissue source	Clinical applications or treatment
Preclinical	Central nervous system	Stroke, spinal-cord injury
	Muscle	Volumetric muscle loss
	Lung, liver, heart, kidney	Organ replacement
Clinical testing	Heart	Myocardial infarction
	Adipose	Soft-tissue reconstruction
	Small intestine	Volumetric muscle loss, congenital heart-defect repair
	Urinary bladder	Volumetric muscle loss, oesophageal repair
Commercial	Bone	Bone graft
	Skin	Hernia repair, breast reconstruction, dural repair
	Small intestine	Hernia repair, orthopaedic reconstruction, cardiac repair
	Pericardium	Hernia repair, dural repair
	Urinary bladder	Hernia repair, diabetic ulcer, wound healing
	Heart valve	Heart-valve replacement

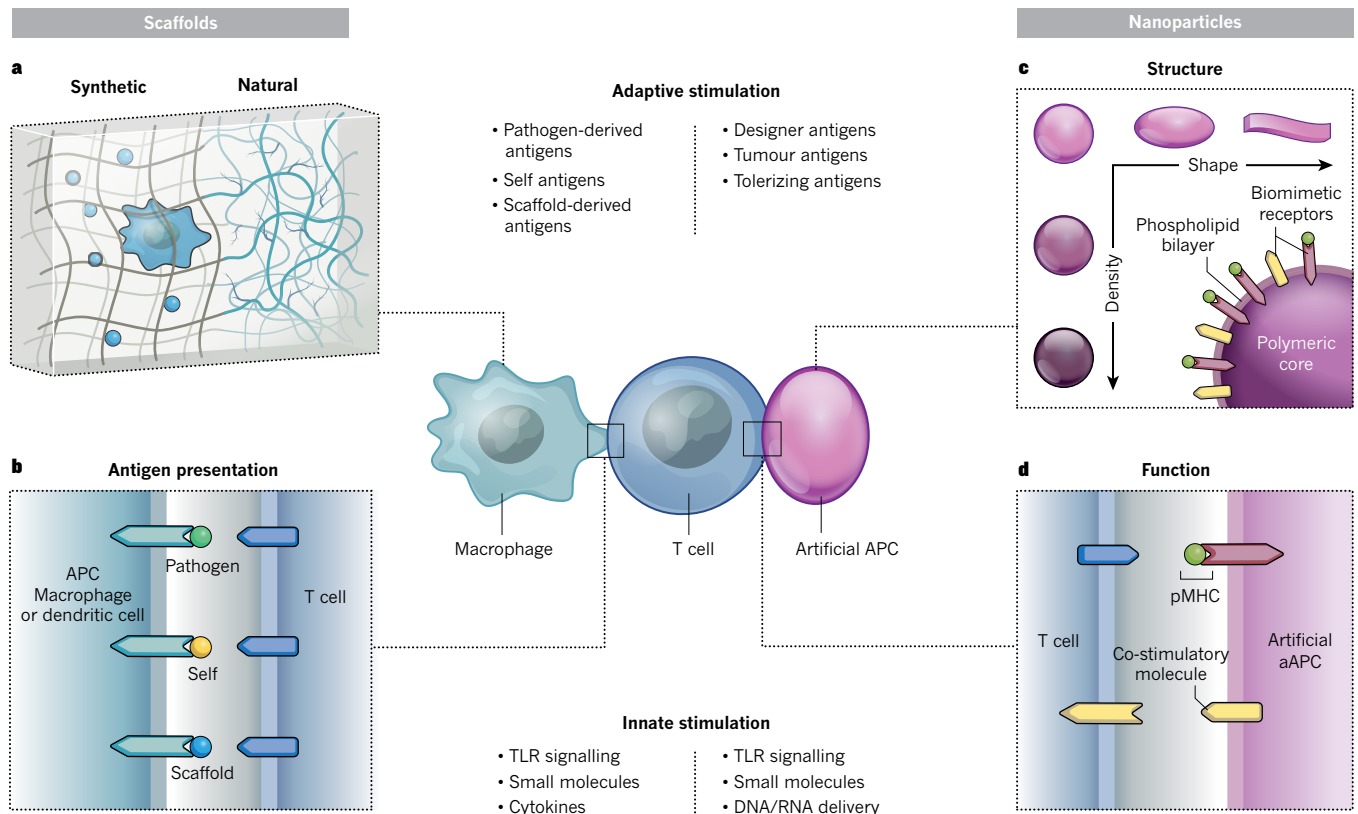


Figure 4 | Synthetic polymer structures used for active interaction with immune cells. **a,b**, Scaffolds. Immune cells trapped or migrating into scaffolds can be manipulated by signals embedded in the material (**a**). Scaffolds can also include elements that are processed by antigen-presenting cells (APCs) (**b**) and presented to T cells where they can induce the desired response. **c,d**, Particles.

biodegradable elastomer poly(glycerol-co-sebacate acrylate) has been formed into pillars like those of gecko feet (see Fig. 3a) and coated with oxidized dextran, which reacts with biological tissue. This combination of mechanical interlocking and covalent integration has been used to treat abdominal wounds and has broad potential in wound healing⁶⁹. The swelling properties of hydrogels can be used in pillars like sticky tape to lock a material into some tissue, in a similar way to how endoparasitic worms anchor themselves in the wet environment of the gut⁷⁰.

Similar bio-inspiration can also be used to achieve the opposite result: a slippery coating that does not adhere to biological materials and so can be used to avoid platelet adhesion, clotting and bio-fouling (Fig. 3d). Mimicking the slippery surface properties of carnivorous plants that use a molecularly smooth liquid surface layer to entrap insects, medical implants have been given a bilayer of tethered and mobile perfluorodecalin to prevent bio-fouling and clot formation in a pig arteriovenous shunt²⁷. Combining this slippery nanocoating with the geometry of surface bumps found on desert beetles and cactus can lead to surface functions such as skin moisturizing²⁹. Moving the concept of slippery surfaces to the body, tissues such as cartilage in articulating joints and lungs require a slippery surface. Interactions between tissue surfaces and surrounding fluids in these spaces create a low-friction environment. Mimicking these tissue-fluid interfaces with synthetic materials³¹, the development of synthetic 'hairy brushes', based on a methacrylate backbone and polyzwitterionic branches or a poly(*n*-isopropylacrylamide) backbone, provides another therapeutic arena for biomimetic materials^{71,72} (Fig. 3e). These synthetic strategies mimic aspects of biological structure and performance but are purely synthetic in chemical composition, facilitating reproducible production and manufacturing scale up.

Combining synthetic and biological approaches, researchers have replicated the function of the cartilage surface protein lubricin³¹,

which binds hyaluronic acid (HA) and is lost in arthritic diseases. An artificial HA-binding peptide sequence was connected to a linker and a collagen-binding peptide for attachment to tissues after injection. In contrast to current clinical therapies of supplementing HA in the joint, this technique uses the endogenous HA already present in the joint space and binds it to cartilage to enhance surface lubrication and partly rebuild the role of lost or damaged surface protein.

Future perspectives

The biocompatibility of materials has historically focused on reducing biological interactions to minimize the risk of rejection for non-cellular biomaterials such as pacemakers. But now we demand that materials have active biological functions such as sensing or stimulating elements of the local environment. The immune system in particular is becoming a focus for manipulation. Using biomaterials to target specific components of the immune system, in multiple forms and in combination with drug delivery, has the potential to significantly increase the ability of already game-changing cancer immunotherapies to attack and kill tumours, for example. Implants made of biomaterials, on the other hand, are moving towards directing the immune system to avoid the foreign-body response and fibrosis. Beyond cancer immunotherapy and the biocompatibility of devices, the role of the immune system in repairing and regenerating tissues highlights the potentially broad use of immunomodulatory materials to stimulate tissue growth^{21,73} (Fig. 4). The capabilities of biomaterials in all these areas of biomedical research will develop exponentially if we can move towards an in-depth, mechanistic understanding of how they interact with the immune system both locally and systemically.

The fundamental biology of tissue structures (such as the proteomic composition of the ECM) and the details of the immune response to materials are still being studied, highlighting again the need for

biologists and engineers to work together to move forward both the basic science and the therapeutic development of materials. An important component of the future success of polymers in biomedical applications is a better understanding of what factors can predict success. Systematic evaluation of which interactions of biomaterials with cells or tissues correlate with improved performance in physiological environments is just beginning. Two recent studies showed how combinatorial chemical modifications of alginate hydrogels could be evaluated in mice and monkeys, with the resulting 'optimal' material being used to encapsulate insulin-producing human stem-cell-derived β cells for transplantation in mice^{28–31}. Even in ECM materials that have a complex composition, high-throughput screening of cells' responses to different tissue ECMs, in combination with proteomic analysis and systems-biology techniques, has defined new potential mechanisms of cellular responses⁷⁴. Unfortunately, the challenge of predicting the responses to materials in clinical applications becomes even more complex when one considers that they are usually applied therapeutically in the diseased or otherwise abnormal environments typical of the patients that most need them⁷⁵.

Received 6 January; accepted 12 September 2016.

1. Rosales, A. M. & Anseth, K. S. The design of reversible hydrogels to capture extracellular matrix dynamics. *Nature Rev. Mater.* **1**, 15012 (2016).
2. Engler, A. J., Sen, S., Sweeney, H. L. & Discher, D. E. Matrix elasticity directs stem cell lineage specification. *Cell* **126**, 677–689 (2006).
This is a landmark paper on the role of a hydrogel scaffold's mechanical properties in stem-cell differentiation.
3. Nelson, C. M., VanDuijn, M. M., Inman, J. L., Fletcher, D. A. & Bissell, M. J. Tissue geometry determines sites of mammary branching morphogenesis in organotypic cultures. *Science* **314**, 298–300 (2006).
4. Ye, C. *et al.* Self-(un)rolling biopolymer microstructures: Rings, tubules, and helical tubules from the same material. *Angew. Chem. Int. Edn Engl.* **54**, 8490–8493 (2015).
5. Benoit, D. S. W., Schwartz, M. P., Durney, A. R. & Anseth, K. S. Small functional groups for controlled differentiation of hydrogel-encapsulated human mesenchymal stem cells. *Nature Mater.* **7**, 816–823 (2008).
6. Lin, C. C. & Anseth, K. S. Cell-cell communication mimicry with poly(ethylene glycol) hydrogels for enhancing beta-cell function. *Proc. Natl Acad. Sci. USA* **108**, 6380–6385 (2011).
7. Lutolf, M. P. & Hubbell, J. A. Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. *Nature Biotechnol.* **23**, 47–55 (2005).
Pioneering study describing how synthetic hydrogels can be used to mimic the native extracellular matrix.
8. Lee, T. T. *et al.* Light-triggered in vivo activation of adhesive peptides regulates cell adhesion, inflammation and vascularization of biomaterials. *Nature Mater.* **14**, 352–360 (2015).
9. Martino, M. M. *et al.* Growth factors engineered for super-affinity to the extracellular matrix enhance tissue healing. *Science* **343**, 885–888 (2014).
10. DeForest, C. A. & Tirrell, D. A. A photoreversible protein-patterning approach for guiding stem cell fate in three-dimensional gels. *Nature Mater.* **14**, 523–531 (2015).
11. Chaudhuri, O. *et al.* Hydrogels with tunable stress relaxation regulate stem cell fate and activity. *Nature Mater.* **15**, 326–334 (2016).
12. Dingal, P. C. D. P. *et al.* Fractal heterogeneity in minimal matrix models of scars modulates stiff-niche stem-cell responses via nuclear exit of a mechanorepressor. *Nature Mater.* **14**, 951–960 (2015).
13. Beck, J. N., Singh, A., Rothenberg, A. R., Elisseff, J. H. & Ewald, A. J. The independent roles of mechanical, structural and adhesion characteristics of 3D hydrogels on the regulation of cancer invasion and dissemination. *Biomaterials* **34**, 9486–9495 (2013).
14. Chaudhuri, O. *et al.* Extracellular matrix stiffness and composition jointly regulate the induction of malignant phenotypes in mammary epithelium. *Nature Mater.* **13**, 970–978 (2014).
15. Johnson, R. & Halder, G. The two faces of Hippo: targeting the Hippo pathway for regenerative medicine and cancer treatment. *Nature Rev. Drug Discov.* **13**, 63–79 (2014).
16. Sun, Y. *et al.* Hippo/YAP-mediated rigidity-dependent motor neuron differentiation of human pluripotent stem cells. *Nature Mater.* **13**, 599–604 (2014).
17. Sur, S., Matson, J. B., Webber, M. J., Newcomb, C. J. & Stupp, S. I. Photodynamic control of bioactivity in a nanofiber matrix. *ACS Nano* **6**, 10776–10785 (2012).
18. Kloxin, A. M., Kasko, A. M., Salinas, C. N. & Anseth, K. S. Photodegradable hydrogels for dynamic tuning of physical and chemical properties. *Science* **324**, 59–63 (2009).
19. Gilbert, T. W., Sellaro, T. L. & Badyal, S. F. Decellularization of tissues and organs. *Biomaterials* **27**, 3675–3683 (2006).
20. Badyal, S. F. & Gilbert, T. W. Immune response to biologic scaffold materials. *Semin. Immunol.* **20**, 109–116 (2008).
21. Sadtler, K. *et al.* Developing a pro-regenerative biomaterial scaffold microenvironment requires T helper 2 cells. *Science* **352**, 366–370 (2016).
22. Ali, O. A., Tayalia, P., Shvartsman, D., Lewin, S. & Mooney, D. J. Inflammatory cytokines presented from polymer matrices differentially generate and activate DCs. *Adv. Funct. Mater.* **23**, 4621–4628 (2013).
This is the first study to incorporate immunological cytokines into a biomaterial scaffold to modulate an immune response.
23. Kim, J. *et al.* Injectable, spontaneously assembling, inorganic scaffolds modulate immune cells *in vivo* and increase vaccine efficacy. *Nature Biotechnol.* **33**, 64–72 (2015).
24. Ahn, B. K., Lee, D. W., Israelachvili, J. N. & Waite, J. H. Surface-initiated self-healing of polymers in aqueous media. *Nature Mater.* **13**, 867–872 (2014).
25. Damo, M., Wilson, D. S., Simeoni, E. & Hubbell, J. A. TLR-3 stimulation improves anti-tumor immunity elicited by dendritic cell exosome-based vaccines in a murine model of melanoma. *Sci. Rep.* **5**, 17622 (2015).
26. Veisoh, O. *et al.* Size- and shape-dependent foreign body immune response to materials implanted in rodents and non-human primates. *Nature Mater.* **14**, 643–651 (2015).
27. Leslie, D. C. *et al.* A bioinspired omniphobic surface coating on medical devices prevents thrombosis and biofouling. *Nature Biotechnol.* **32**, 1134–1140 (2014).
This paper introduces surface topographies from nature to control biomaterial surface properties.
28. Vegas, A. J. *et al.* Combinatorial hydrogel library enables identification of materials that mitigate the foreign body response in primates. *Nature Biotechnol.* **34**, 345–352 (2016).
This pioneering study demonstrates that a combinatorial library approach of constructing synthetic alginate variants can lead to biomaterials that reduce foreign-body reactions in non-human primates for at least 6 months.
29. Park, K.-C. *et al.* Condensation on slippery asymmetric bumps. *Nature* **531**, 78–82 (2016).
30. Vegas, A. J. *et al.* Long-term glycemic control using polymer-encapsulated human stem cell-derived beta cells in immune-competent mice. *Nature Med.* **22**, 306–311 (2016).
31. Singh, A. *et al.* Enhanced lubrication on tissue and biomaterial surfaces through peptide-mediated binding of hyaluronic acid. *Nature Mater.* **13**, 988–995 (2014).
32. Kelich, J. M. *et al.* Super-resolution imaging of nuclear import of adeno-associated virus in live cells. *Mol. Ther. Meth. Clin. Dev.* **2**, 15047 (2015).
33. Goldsmith, C. S. & Miller, S. E. Modern uses of electron microscopy for detection of viruses. *Clin. Microbiol. Rev.* **22**, 552–563 (2009).
34. Sachse, C. *et al.* High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *J. Mol. Biol.* **371**, 812–835 (2007).
35. Vestergaard, G. *et al.* Stygiolobus rod-shaped virus and the interplay of crenarchaeal rudiaviruses with the CRISPR antiviral system. *J. Bacteriol.* **190**, 6837–6845 (2008).
36. Bharat, T. A. *et al.* Structural dissection of Ebola virus and its assembly determinants using cryo-electron tomography. *Proc. Natl Acad. Sci. USA* **109**, 4275–4280 (2012).
37. Häring, M. *et al.* Virology: independent virus development outside a host. *Nature* **436**, 1101–1102 (2005).
38. Jiang, X. *et al.* Plasmid-templated shape control of condensed DNA-block copolymer nanoparticles. *Adv. Mater.* **25**, 227–232 (2013).
This paper establishes that the shape of polymeric, plasmid DNA-containing nanoparticles can be controlled by solvent polarity, and that anisotropic biomimetic particles can have enhanced gene-delivery efficacy *in vivo*.
39. Hanson, M. C., Bershteyn, A., Crespo, M. P. & Irvine, D. J. Antigen delivery by lipid-enveloped PLGA microparticle vaccines mediated by *in situ* vesicle shedding. *Biomacromolecules* **15**, 2475–2481 (2014).
40. Roberts, R. A. *et al.* Towards programming immune tolerance through geometric manipulation of phosphatidylserine. *Biomaterials* **72**, 1–10 (2015).
41. Perry, J. L., Herlihy, K. P., Napier, M. E. & Desimone, J. M. PRINT: a novel platform toward shape and size specific nanoparticle therapeutics. *Acc. Chem. Res.* **44**, 990–998 (2011).
42. Hu, C. M. *et al.* Erythrocyte membrane-camouflaged polymeric nanoparticles as a biomimetic delivery platform. *Proc. Natl Acad. Sci. USA* **108**, 10980–10985 (2011).
This paper uses cell membranes to camouflage and functionalize polymeric nanoparticles, opening the door to new hybrid biomimetic particles.
43. Fang, R. H. *et al.* Cancer cell membrane-coated nanoparticles for anticancer vaccination and drug delivery. *Nano Lett.* **14**, 2181–2188 (2014).
44. Hu, C.-M. J. *et al.* Nanoparticle biointerfacing by platelet membrane cloaking. *Nature* **526**, 118–121 (2015).
45. Hu, C.-M. J., Fang, R. H., Copp, J., Luk, B. T. & Zhang, L. A biomimetic nanosponge that absorbs pore-forming toxins. *Nature Nanotechnol.* **8**, 336–340 (2013).
46. Parodi, A. *et al.* Synthetic nanoparticles functionalized with biomimetic leukocyte membranes possess cell-like functions. *Nature Nanotechnol.* **8**, 61–68 (2013); published online 16 December 2012.
47. Xuan, M., Shao, J., Dai, L., He, Q. & Li, J. Macrophage cell membrane camouflage mesoporous silica nanocapsules for *in vivo* cancer therapy. *Adv. Healthcare Mater.* **4**, 1645–1652 (2015).
48. Lai, P.-Y., Huang, R.-Y., Lin, S.-Y., Lin, Y.-H. & Chang, C.-W. Biomimetic stem cell membrane-camouflaged iron oxide nanoparticles for theranostic applications. *RSC Adv.* **5**, 98222–98230 (2015).
49. Rodriguez, P. L. *et al.* Minimal 'self' peptides that inhibit phagocytic clearance and enhance delivery of nanoparticles. *Science* **339**, 971–975 (2013).

50. Tsai, R. K., Rodriguez, P. L. & Discher, D. E. Self inhibition of phagocytosis: the affinity of 'marker of self' CD47 for SIRPalpha dictates potency of inhibition but only at low expression levels. *Blood Cells Mol. Dis.* **45**, 67–74 (2010).
 51. Sosale, N. G. *et al.* Cell rigidity and shape override CD47's "self"-signaling in phagocytosis by hyperactivating myosin-II. *Blood* **125**, 542–552 (2015).
 52. Wilhelm, S. *et al.* Analysis of nanoparticle delivery to tumours. *Nature Rev. Mater.* **1**, 16014 (2016).
 53. Meyer, R. A. *et al.* Biodegradable nanoellipsoidal artificial antigen presenting cells for antigen specific T-cell activation. *Small* **11**, 1519–1525 (2015).
 54. Perica, K. *et al.* Enrichment and expansion with nanoscale artificial antigen presenting cells for adoptive immunotherapy. *ACS Nano* **9**, 6861–6871 (2015).
 55. Sunshine, J. C., Perica, K., Schneck, J. P. & Green, J. J. Particle shape dependence of CD8+ T cell activation by artificial antigen presenting cells. *Biomaterials* **35**, 269–277 (2014).
 56. Lashof-Sullivan, M. M. *et al.* Intravenously administered nanoparticles increase survival following blast trauma. *Proc. Natl Acad. Sci. USA* **111**, 10293–10298 (2014).
 57. Anselmo, A. C. *et al.* Platelet-like nanoparticles: mimicking shape, flexibility, and surface biology of platelets to target vascular injuries. *ACS Nano* **8**, 11243–11253 (2014).
 58. Chan, L. W. *et al.* A synthetic fibrin cross-linking polymer for modulating clot properties and inducing hemostasis. *Sci. Transl. Med.* **7**, 277ra29 (2015).
 59. Kumar, V. A., Wickremasinghe, N. C., Shi, S. & Hartgerink, J. D. Nanofibrous snake venom hemostat. *ACS Biomater. Sci. Eng.* **1**, 1300–1305 (2015).
 60. Lee, H., Dellatore, S. M., Miller, W. M. & Messersmith, P. B. Mussel-inspired surface chemistry for multifunctional coatings. *Science* **318**, 426–430 (2007).
 61. Maier, G. P., Rapp, M. V., Waite, J. H., Israelachvili, J. N. & Butler, A. Adaptive synergy between catechol and lysine promotes wet adhesion by surface salt displacement. *Science* **349**, 628–632 (2015).
 62. Papanna, R. *et al.* Cryopreserved human amniotic membrane and a bioinspired underwater adhesive to seal and promote healing of iatrogenic fetal membrane defect sites. *Placenta* **36**, 888–894 (2015).
 63. Zhao, Q. *et al.* Underwater contact adhesion and microarchitecture in polyelectrolyte complexes actuated by solvent exchange. *Nature Mater.* **15**, 407–412 (2016).
 64. Lee, Y. *et al.* Bioinspired nanoparticulate medical glues for minimally invasive tissue repair. *Adv. Healthcare Mater.* **4**, 2587–2596 (2015).
 65. Roche, E. T. *et al.* A light-reflecting balloon catheter for atraumatic tissue defect repair. *Sci. Transl. Med.* **7**, 306ra149 (2015).
 66. Busscher, H. J. *et al.* Biomaterial-associated infection: Locating the finish line in the race for the surface. *Sci. Transl. Med.* **4**, 153rv10 (2012).
 67. Geim, A. K. *et al.* Microfabricated adhesive mimicking gecko foot-hair. *Nature Mater.* **2**, 461–463 (2003).
 68. Lee, H., Lee, B. P. & Messersmith, P. B. A reversible wet/dry adhesive inspired by mussels and geckos. *Nature* **448**, 338–341 (2007).
 69. Mahdavi, A. *et al.* A biodegradable and biocompatible gecko-inspired tissue adhesive. *Proc. Natl Acad. Sci. USA* **105**, 2307–2312 (2008).
 70. Yang, S. Y. *et al.* A bio-inspired swellable microneedle adhesive for mechanical interlocking with tissue. *Nature Commun.* **4**, 1702 (2013).
 71. Chen, M., Briscoe, W. H., Armes, S. P. & Klein, J. Lubrication at physiological pressures by polyzwitterionic brushes. *Science* **323**, 1698–1701 (2009).
 72. Liu, G. *et al.* Hairy polyelectrolyte brushes-grafted thermosensitive microgels as artificial synovial fluid for simultaneous biomimetic lubrication and arthritis treatment. *ACS Appl. Mater. Interfaces* **6**, 20452–20463 (2014).
 73. Sicari, B. M. *et al.* An acellular biologic scaffold promotes skeletal muscle formation in mice and humans with volumetric muscle loss. *Sci. Transl. Med.* **6**, 234ra58 (2014).
 74. Beachley, V. Z. *et al.* Tissue matrix arrays for high-throughput screening and systems analysis of cell function. *Nature Methods* **12**, 1197–1204 (2015).
 75. Oliva, N. *et al.* Regulation of dendrimer/dextran material performance by altered tissue microenvironment in inflammation and neoplasia. *Sci. Transl. Med.* **7**, 272ra11 (2015).
- This paper introduces the challenge posed by diverse physiological environments and shows they affect the responses of biomaterials in people.**
76. Banquy, X., Burdyńska, J., Lee, D. W., Matyjaszewski, K. & Israelachvili, J. Bioinspired bottle-brush polymer exhibits low friction and amontons-like behavior. *J. Am. Chem. Soc.* **136**, 6199–6202 (2014).

Acknowledgements The authors thank K. Sadtler for contributions and the design of Figs 1 and 4, C. Cherry for editorial assistance, and M. Frisk for critical review and manuscript contributions. J.J.G. was supported in part by the NIH (1R01EB016721). J.H.E. was supported by the Department of Defense including the Armed Forces Institute of Regenerative Medicine.

Author contributions J.J.G. and J.H.E. both contributed equally to the planning and writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/2fs8s23. Correspondence should be addressed to J.H.E. (jhe@jhu.edu).

The seahorse genome and the evolution of its specialized morphology

Qiang Lin^{1*§}, Shaohua Fan^{2†*}, Yanhong Zhang^{1*}, Meng Xu^{3*}, Huixian Zhang^{1,4*}, Yulan Yang^{3*}, Alison P. Lee^{4†}, Joost M. Woltering², Vydiathan Ravi⁴, Helen M. Gunter^{2†}, Wei Luo¹, Zexia Gao⁵, Zhi Wei Lim^{4†}, Geng Qin^{1,6}, Ralf F. Schneider², Xin Wang^{1,6}, Peiwen Xiong², Gang Li¹, Kai Wang⁷, Jiumeng Min³, Chi Zhang³, Ying Qiu⁸, Jie Bai⁸, Weiming He³, Chao Bian⁸, Xinhui Zhang⁸, Dai Shan³, Hongyue Qu^{1,6}, Ying Sun⁸, Qiang Gao³, Liangmin Huang^{1,6}, Qiong Shi^{1,8§}, Axel Meyer^{2§} & Byrappa Venkatesh^{4,9§}

Seahorses have a specialized morphology that includes a toothless tubular mouth, a body covered with bony plates, a male brood pouch, and the absence of caudal and pelvic fins. Here we report the sequencing and *de novo* assembly of the genome of the tiger tail seahorse, *Hippocampus comes*. Comparative genomic analysis identifies higher protein and nucleotide evolutionary rates in *H. comes* compared with other teleost fish genomes. We identified an astacin metalloprotease gene family that has undergone expansion and is highly expressed in the male brood pouch. We also find that the *H. comes* genome lacks enamel matrix protein-coding proline/glutamine-rich secretory calcium-binding phosphoprotein genes, which might have led to the loss of mineralized teeth. *tbx4*, a regulator of hindlimb development, is also not found in *H. comes* genome. Knockout of *tbx4* in zebrafish showed a 'pelvic fin-loss' phenotype similar to that of seahorses.

Members of the teleost family Syngnathidae (seahorses, pipefishes and seadragons) (Extended Data Fig. 1), comprising approximately 300 species, display a complex array of morphological innovations and reproductive behaviours. This includes specialized morphological phenotypes such as an elongated snout with a small terminal mouth, fused jaws, absent pelvic and caudal fins, and an extended body covered with an armour of bony plates instead of scales¹ (Fig. 1a). Syngnathids are also unique among vertebrates due to their 'male pregnancy', whereby males nourish developing embryos in a brood pouch until hatching and parturition occurs^{2,3}. In addition, members of the sub-family Hippocampinae (seahorses) exhibit other derived features such as the lack of a caudal fin, a characteristic prehensile tail, and a vertical body axis⁴ (Fig. 1a). To understand the genetic basis of the specialized morphology and reproductive system of seahorses, we sequenced the genome of the tiger tail seahorse, *H. comes*, and carried out comparative genomic analyses with the genome sequences of other ray-finned fishes (Actinopterygii).

Genome assembly and annotation

The genome of a male *H. comes* individual was sequenced using the Illumina HiSeq 2000 platform. After filtering low-quality and duplicate reads, 132.13 Gb (approximately 190-fold coverage of the estimated 695 Mb genome) of reads from libraries with insert sizes ranging from 170 bp to 20 kb were retained for assembly. The filtered reads were assembled using SOAPdenovo (version 2.04) to yield a 501.6 Mb assembly with an N50 contig size and N50 scaffold size of 34.7 kb and 1.8 Mb, respectively. Total RNA from combined soft tissues of *H. comes* was sequenced using RNA-sequencing (RNA-seq) and assembled

de novo. The *H. comes* genome assembly is of high quality, as >99% of the *de novo* assembled transcripts (76,757 out of 77,040) could be mapped to the assembly; and 243 out of 248 core eukaryotic genes mapping approach (CEGMA) genes are complete in the assembly.

We predicted 23,458 genes in the genome of *H. comes* based on homology and by mapping the RNA-seq data of *H. comes* and a closely related species, the lined seahorse, *Hippocampus erectus*, to the genome assembly (see Methods and Supplementary Information). More than 97% of the predicted genes (22,941 genes) either have homologues in public databases (Swissprot, TrEMBL and the Kyoto Encyclopedia of Genes and Genomes (KEGG)) or are supported by assembled RNA-seq transcripts. Analysis of gene family evolution using a maximum likelihood framework identified an expansion of 25 gene families (261 genes; 1.11%) and contraction of 54 families (96 genes; 0.41%) in the *H. comes* lineage (Extended Data Fig. 2 and Supplementary Tables 4.1, 4.2). Transposable elements comprise around 24.8% (124.5 Mb) of the *H. comes* genome, with class II DNA transposons being the most abundant class (9%; 45 Mb). Only one wave of transposable element expansion was identified, with no evidence for a recent transposable element burst (Kimura divergence ≤ 5) (Extended Data Fig. 3).

Phylogenomics and evolutionary rate

The phylogenetic relationships between *H. comes* and other teleosts were determined using a genome-wide set of 4,122 one-to-one orthologous genes (Supplementary Note 4.2). The phylogenetic analysis (Fig. 1b) showed that *H. comes* is a sister group to other percomorph fishes analysed (stickleback, *Gasterosteus aculeatus*; medaka, *Oryzias latipes*; Nile tilapia, *Oreochromis niloticus*; fugu, *Takifugu rubripes*; and

¹CAS Key Laboratory of Tropical Marine Bio-resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China. ²Chair in Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, Konstanz 78457, Germany. ³BGI-Shenzhen, Shenzhen 518083, China. ⁴Institute of Molecular and Cell Biology, A*STAR, Biopolis, Singapore 138673, Singapore. ⁵College of Fisheries, Huazhong Agricultural University, Wuhan 430070, China. ⁶University of Chinese Academy of Science, Beijing 100049, China. ⁷School of Agriculture, Ludong University, Yantai 264025, China. ⁸Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI, Shenzhen 518083, China. ⁹Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore. [†]Present addresses: Department of Genetics, University of Pennsylvania, Pennsylvania 19104, USA (S.F.); Bioprocessing Technology Institute, Biopolis, Singapore 138668, Singapore (A.P.L.); Institute of Evolutionary Biology, the University of Edinburgh EH9 3FL, UK (H.M.G.); School of Material Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore (Z.W.L.).

*These authors contributed equally to this work.

§These authors jointly supervised this work.

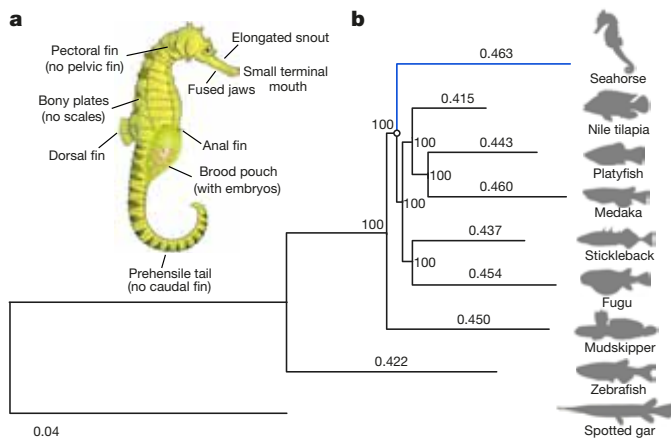


Figure 1 | Adaptations and evolutionary rate of *H. comes*. **a**, Schematic diagram of a pregnant male seahorse. **b**, The phylogenetic tree generated using protein sequences. The values on the branches are the distances (number of substitutions per site) between each of the teleost fishes and the spotted gar (outgroup). Spotted gar, *Lepisosteus oculatus*; zebrafish, *Danio rerio*.

platyfish, *Xiphophorus maculatus*) with the exception of blue-spotted mudskipper (*Boleophthalmus pectinirostris*), a member of the family Gobiidae. Our inference, which placed the mudskipper as the outgroup, differs from that of a previous phylogenetic analysis based on fewer protein-coding genes that had placed syngnathids as an outgroup⁵. Estimated divergence times of *H. comes* and other teleosts calculated using MCMCTree suggest that *H. comes* diverged from the other percomorphs approximately 103.8 million years ago, during the Cretaceous period (Extended Data Fig. 2). Interestingly, the branch length of *H. comes* is longer than that of other teleosts, suggesting a higher protein evolutionary rate compared to other teleosts analysed in this study (Fig. 1b). This result was found to be statistically significant by both relative rate test⁶ and two cluster analysis⁷ (Supplementary Tables 4.3 and 4.4). To determine whether the neutral nucleotide substitution rate of *H. comes* is also higher, we generated a neutral tree on the basis of fourfold degenerate sites and calculated the pairwise distance of each teleost to the spotted gar (an outgroup) (Supplementary Fig. 4.4). The pairwise distance of *H. comes* was again higher compared with other teleosts, indicating that the neutral evolutionary rate of *H. comes* is also higher than that of other teleosts. The reasons for this higher molecular evolutionary rate in *H. comes* are unclear.

Gene loss

Gene loss or loss of function can contribute to evolutionary novelties and can be positively selected for^{8,9}. We identified several genes that are not found in the *H. comes* genome but are found in other sequenced teleost genomes.

Secretory calcium-binding phosphoprotein (SCPP) genes encode extracellular matrix proteins that are involved in the formation of mineralized tissues such as bone, dentin, enamel and enameloid. Bony vertebrate genomes encode multiple SCPP genes that can be divided into two groups, the acidic and the proline/glutamine (P/Q)-rich SCPP genes. Acidic SCPPs regulate the mineralization of collagen scaffolds in bone and dentin whereas the P/Q-rich SCPPs are primarily involved in enamel or enameloid formation¹⁰. Analysis of the *H. comes* genome and the transcriptomes of *H. comes* and *H. erectus* showed that both contain two acidic SCPP genes, *scpp1* and *spp1* (Extended Data Fig. 4). However, no intact P/Q-rich gene could be identified. The only P/Q-rich gene present in the *H. comes* genome assembly, *scpp5*, is represented by only three out of ten exons, indicating that it has become a pseudogene. Seahorses and pipefish (family Syngnathidae) are toothless, a phenomenon known as edentulism. Besides syngnathids, edentulism has occurred convergently in several other vertebrate

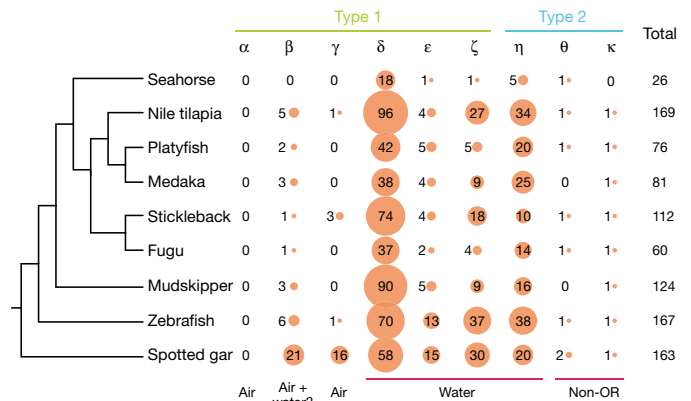


Figure 2 | OR genes in *H. comes* and other ray-finned fishes. ‘Air’ and ‘water’ refer to the detection of airborne and water-soluble odorants, respectively. The sizes of the orange circles represent the number of OR genes of a particular category.

lineages¹¹, the most notable ones being birds¹², turtles, and some mammals such as baleen whales, pangolins and anteaters¹³. The loss of teeth in birds, turtles and mammals has been attributed to inactivating mutations in one or more P/Q-rich enamel-specific SCPP genes such as *Enam*, *Amel*, *Ambn* and *Amtm*, and the dentin-specific gene, *Dspp*^{12,14}. In the case of *H. comes*, the complete loss of functional P/Q-rich SCPP genes may explain the loss of mineralized teeth.

Animals use their sense of smell, or olfaction, for finding food, mates and avoiding predators. Olfaction is mediated by olfactory receptors (ORs), which constitute the largest family of G-protein-coupled receptors. We were able to identify in the *H. comes* genome a significantly smaller repertoire of OR genes than in other teleosts (*P* value < 0.05, Wilcoxon rank-sum test). Our sensitive search pipeline (based on TblastN and Genewise) and manual inspection identified only 26 OR genes in the *H. comes* genome—the smallest OR repertoire identified in any ray-finned fish genome analysed so far (60 to 169 OR genes) (Fig. 2 and Extended Data Fig. 5).

A derived phenotype of seahorse and other syngnathids is the complete lack of pelvic fins^{15,16}. Pelvic fins are homologous to tetrapod hindlimbs and primarily serve a role in body trim and subtle swimming manoeuvres during teleost locomotion^{17–19}. In addition, pelvic spines have an important role in protection against predators¹⁵. Pelvic fin loss has occurred independently in several teleost lineages, including Tetraodontidae (for example, pufferfishes), Anguillidae (eels) and Gasterosteidae (some populations of sticklebacks), and is frequently associated with a reduced pressure from predators and/or the evolution of an elongated body plan¹⁵. In pufferfish (fugu), pelvic fin loss is associated with a change in the expression pattern of *hoxd9a*²⁰. In freshwater populations of stickleback, the loss of pelvic fins has been demonstrated to be due to deletions in the pelvic fin-specific enhancer of *pitx1* (ref. 21).

Analysis of the *H. comes* genome and the transcriptomes of *H. comes* and *H. erectus* (see Supplementary Information, section 2), suggested that *tbx4*, a transcription factor conserved in jawed vertebrates, is not present in the seahorse genome (Fig. 3a) (Supplementary Information, section 9). To verify this, we carried out degenerate polymerase chain reaction (PCR) using genomic DNA from *H. comes* and several other species of syngnathids and some non-syngnathids. While the degenerate primers amplified a fragment of *tbx4* from non-syngnathids, they failed to amplify a *tbx4* fragment from syngnathid fishes (see Supplementary Information, section 9). *Tbx4* is a T-box DNA-binding domain-containing transcription factor that acts as a regulator of hindlimb formation in mammals^{22–24}. Loss of function of this gene in mouse leads to a failure of hindlimb formation^{22,23} as well as strong pleiotropic defects in lung²⁵ and placental development²². Expression of zebrafish *tbx4* specifically in pelvic fins suggests a similar role in appendage patterning in fishes²⁴. Given the major role of *tbx4* in

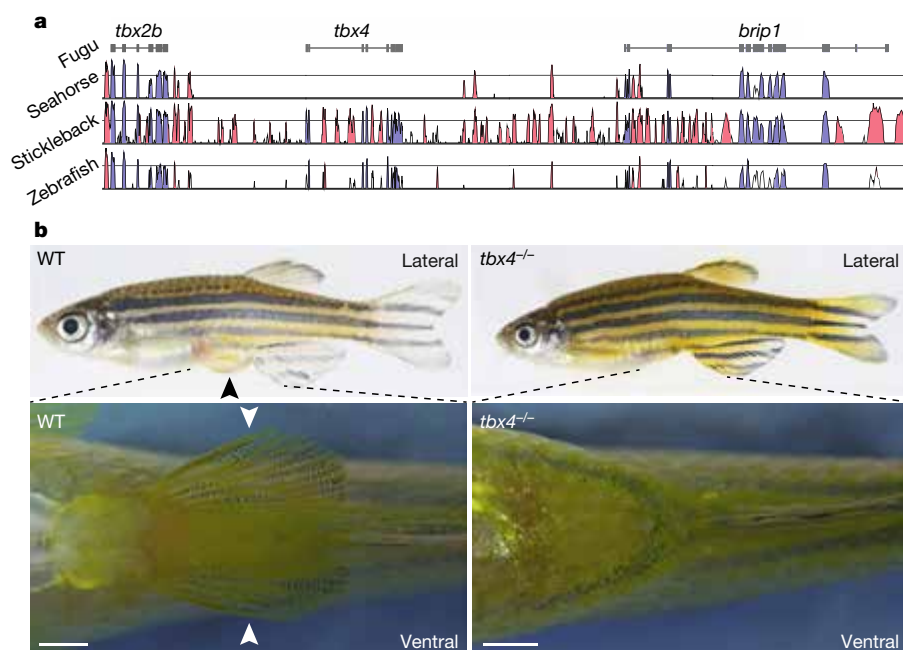


Figure 3 | Pelvic fin loss in *H. comes* is associated with loss of *tbx4*. **a**, Vista plot of conserved elements in the *tbx2b*–*tbx4*–*brip1* syntenic region in fugu (reference genome), seahorse (*H. comes*), stickleback and zebrafish showing that *tbx4* is missing from this locus in seahorse. The blue and red peaks represent conserved exonic and non-coding sequences, respectively. **b**, Lateral (top) and ventral view (bottom) of wild-type (WT) and a representative (one out of five) F3 homozygous *tbx4*-null mutant (*tbx4*^{−/−}) zebrafish. Bottom panel shows a close-up of the pelvic region (dashed lines indicate the approximate zoom region). Scale bar, 1 mm. Pelvic fins are indicated with black or white arrowheads in the wild-type fish. Homozygous *tbx4*-null mutants entirely lack pelvic fins without showing any other gross morphological defects.

hindlimb formation in mammals, we hypothesized that its absence in *H. comes* might be associated with the loss of pelvic fins. To test this hypothesis, we generated a CRISPR–Cas9 *tbx4*-knockout mutant zebrafish line. Interestingly, unlike homozygous mouse *Tbx4* mutants, which fail to develop a functional allantois²², the homozygous zebrafish mutants are viable but completely lack pelvic fins without exhibiting any other gross morphological abnormalities in pectoral or median fins (Fig. 3c and Extended Data Fig. 6; see also Supplementary Information, section 9.3, in particular Supplementary Fig. 9.6 for additional phenotype analysis). This finding is consistent with the results of a recent study that showed that mutations in *tbx4* are associated with the loss of pelvic fins in a naturally occurring zebrafish strain called *pelvic finless*²⁶ (see also Supplementary Information, section 9.3). These results show that *tbx4* has a role in pelvic fin formation in teleosts and suggests that the loss of pelvic fins in *H. comes* may be related to the loss of *tbx4*.

Expansion of the *patristacin* gene family

Male pregnancy is an evolutionary innovation unique to syngnathids. In teleosts, the C6AST subfamily of astacin metalloproteases—such as high choriolytic enzyme (HCE) and low choriolytic enzyme (LCE)—are involved in lysing the chorion surrounding the egg, leading to hatching of embryos²⁷. A member of this subfamily, *patristacin* (*pastn*), was found to be highly expressed in the brood pouch of pregnant males of the Gulf pipefish, *Syngnathus scovelli*, leading to the suggestion that this gene may have a role in the evolution of male pregnancy²⁸. A *pastn* gene was also found to be highly expressed in the brood pouch of the male big belly seahorse, *H. abdominalis*, during mid- and late pregnancy²⁹, suggesting a shared role for this gene in male pregnancy in syngnathids.

The *H. comes* genome contains six *pastn* genes (*pastn1* to *pastn6*; Fig. 4a) organized in a cluster. To examine their expression patterns in the brood pouch, we carried out RNA-seq analysis at different stages of brood pouch development (see Supplementary Information, section 2) in *H. erectus*, as this species is easy to obtain and breed in the laboratory. *H. comes* and *H. erectus* exhibit very similar reproductive cycles and their coding sequences are highly similar (average identity of 93.3%; determined by aligning *H. erectus* RNA-seq transcripts to the *H. comes* genome assembly). We identified orthologues for five of the *H. comes* *pastn* genes (*pastn1*, *pastn2*, *pastn3*, *pastn5* and *pastn6*) in the RNA-seq transcripts of *H. erectus* (Supplementary Fig. 2). Quantitative reverse transcription PCR (qRT–PCR) analysis of these

genes showed that some of them are expressed at significantly higher levels in early- and late-pregnant stages (Fig. 4c). For example, *pastn2* is expressed at significantly higher levels in early- and late-pregnant stages compared to the non-pregnant stage, whereas *pastn1* and *pastn3* are expressed at significantly higher levels during the late-pregnant stage compared to non-pregnant stage (Fig. 4c). This expression pattern suggests a role for these *pastn* genes in brood pouch development and/or hatching of embryos within the brood pouch prior to parturition.

Interestingly, the platyfish (*X. maculatus*), in which fertilization and hatching of eggs occur within the maternal body (ovoviviparity), contains a cluster of six *c6ast* genes (Fig. 4a), with potential hatching enzyme-like activity³⁰. Phylogenetic analysis of *c6ast* family genes in *H. comes*, platyfish and other fishes showed that *H. comes* *pastn* genes and platyfish *c6ast* genes form separate clades (Fig. 4b), indicating that they have expanded independently in the two lineages. Thus, this is an interesting instance of a gene family (C6AST subfamily of astacin metalloproteases) that has undergone expansion independently in different teleost lineages and shows new expression patterns and functions associated with similar evolutionary innovations (that is, ovoviviparity in female platyfish and male pregnancy in seahorse).

Loss of conserved noncoding elements

Vertebrate genomes contain thousands of noncoding elements that are under purifying selection^{31–33}. Many of these conserved noncoding elements (CNEs) function as *cis*-regulatory elements such as enhancers, repressors and insulators^{34,35}. Evolutionary loss of CNEs has important roles in phenotypic differences and morphological innovations^{21,36,37}. To determine the extent of loss of CNEs in seahorse, we predicted genome-wide CNEs in *H. comes* and four other percomorph fishes (stickleback, fugu, medaka and Nile tilapia) using zebrafish as the reference genome (see Supplementary Information). We identified 239,976 CNEs (average size of 168 bp) that are conserved in zebrafish and at least one of the five percomorph fishes (Supplementary Table 6.1). To determine the extent to which CNEs are lost in *H. comes*, we searched for CNEs that are uniquely lost in each of the percomorph fishes. We restricted our analyses to a high-confidence set of CNEs situated in gap-free syntenic intervals (Supplementary Table 6.5). Interestingly, *H. comes* was found to have lost a substantially higher number of CNEs (1,612 CNEs) compared to other percomorphs (fugu, 1,050 CNEs; stickleback, 843 CNEs; medaka, 335 CNEs; Nile tilapia, 281 CNEs) (Supplementary Table 6.6).

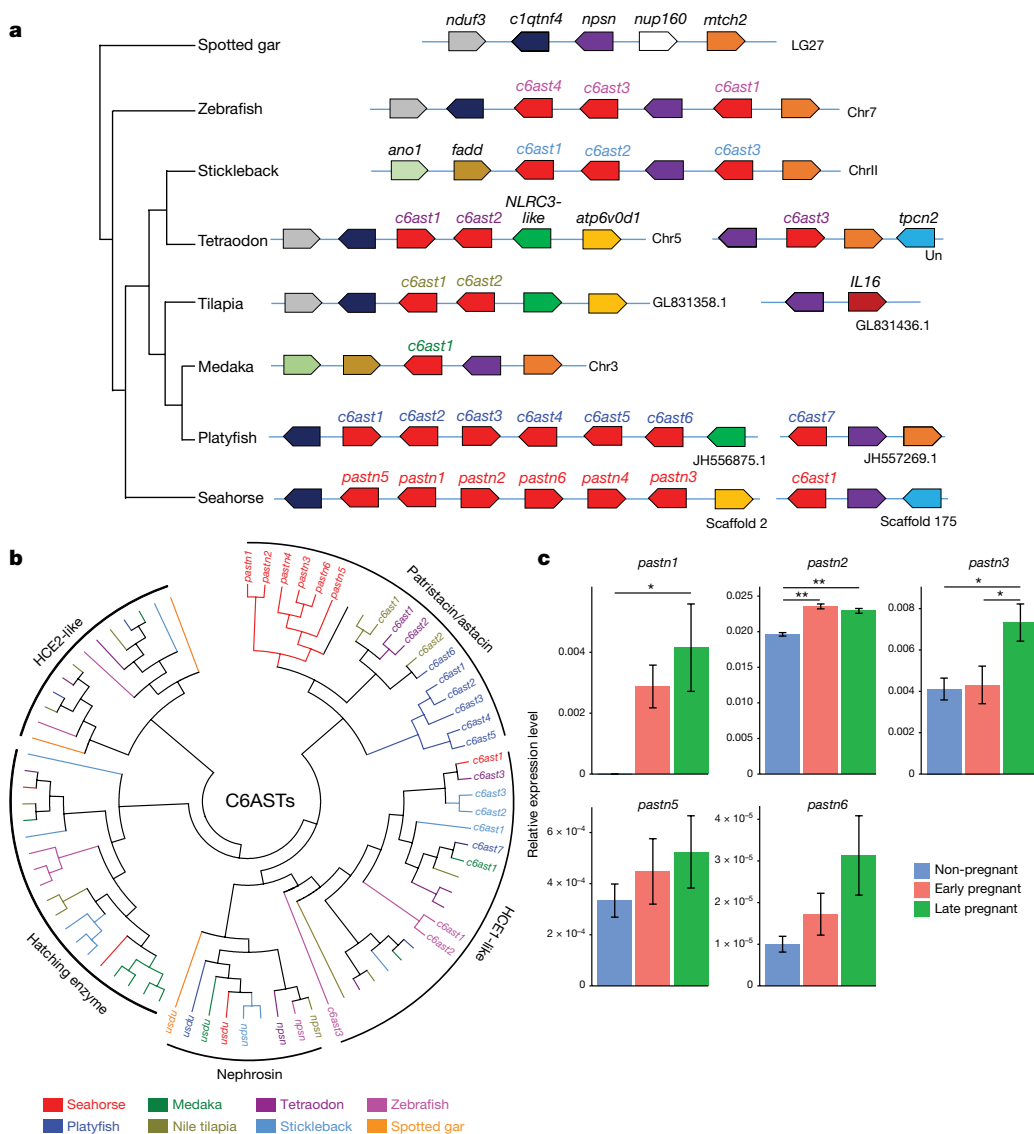


Figure 4 | Astacin metalloproteinase gene family in ray-finned fishes. **a**, Astacin gene loci in various ray-finned fish genomes showing expansion of *pastn* genes in seahorse (*H. comes*) and *c6ast* genes in platyfish. Chr, chromosome. **b**, The phylogeny of the astacin gene family in ray-finned fishes. Only *pastn* or *c6ast* genes shown in **a** are labelled. Supplementary Fig. 10.1 shows an expanded version of the tree with all the genes labelled. **c**, Expression patterns of *pastn* genes in relation to 18S ribosomal RNA genes in the brood pouch of male *H. erectus* determined by qRT-PCR. All data are expressed as mean \pm standard error of mean ($n = 5$) and evaluated by one-way analysis of variance (ANOVA) followed by Tukey's honestly significant difference test for adjusting P values from multiple comparisons (see Methods and Supplementary Information for details of methods). The average duration of pregnancy (from fertilization to parturition) is 17 days⁴¹. The y axis represents expression level in relation to 18S rRNA genes. *pastn1* is expressed at low levels at the non-pregnant stage, which is not clearly visible in the figure due to the large scale used. Non-pregnant: no embryos in the brood pouch; early pregnant: 2–4 days post-fertilization; late pregnant: 12–14 days post-fertilization. * $P < 0.05$, ** $P < 0.01$. Note that *pastn4* is not expressed in these stages of brood pouch.

Analysis of zebrafish CNEs that are lost in *H. comes* indicated that they are present in the neighbourhood of 728 genes enriched in functions such as regulation of transcription, regulation of the fibroblast growth factor receptor signalling pathway, embryonic pectoral fin morphogenesis, steroid hormone receptor activity and O-acetyltransferase activity (Supplementary Tables 6.8 and 6.9). The top 20 genes adjacent to regions with the highest number of CNEs lost in *H. comes* include *sall1a*, *shox* and *irx5a* (Supplementary Tables 6.10 and 6.11), which are involved in the development of the limbs, nervous system, kidney, heart and skeletal system. Altered expression patterns of these genes can potentially lead to altered morphological phenotypes. For example, loss of regulatory regions of the human *SHOX* gene is the cause of Leri–Weill dyschondrosteosis, a dominantly inherited skeletal dysplasia that is characterized by moderate short stature caused by short mesomelic limb segments^{38,39}.

To verify the potential *cis*-regulatory functions of CNEs that were absent in *H. comes* but present in other teleost genomes, we assayed the function of seven selected zebrafish CNEs that were uniquely absent in *H. comes*. Of the seven CNEs assayed in transgenic zebrafish, four CNEs drove reproducible patterns of reporter gene expression in F1 embryos (Extended Data Fig. 7 and Supplementary Table 6.12). Thus, our transgenic assay indicates that some of the CNEs absent in *H. comes* may function as *cis*-regulatory elements in other teleosts. Further studies are required to examine whether the loss of CNEs may have played a role in the evolution of seahorse morphology.

Summary

Seahorses possess one of the most highly specialized morphologies and reproductive behaviours. We sequenced the genome of the tiger tail seahorse and performed comparative analysis with other teleost fishes. Our genome-wide analysis highlights several aspects that may have contributed to the highly specialized body plan and male pregnancy of seahorses. These include a higher protein and nucleotide evolutionary rate, loss of genes and expansion of gene families, with duplicated genes exhibiting new expression patterns, and loss of a selection of potential *cis*-regulatory elements. It is becoming recognized that evolutionary changes in *cis*-regulatory elements, particularly the loss and gain of enhancers, might play a major part in the evolution of morphological innovations and phenotypic changes across species^{21,36,37,40}.

Male pregnancy is a unique developmental feature of seahorses and pipefishes (family Syngnathidae, comprising 57 genera and approximately 300 species). In the seahorse genome, the astacin subfamily of *c6ast* metalloprotease genes has undergone tandem duplications giving rise to six genes. This subfamily of metalloprotease includes the hatching enzyme (also known as choriolyisin), HCE-like and HCE2-like enzymes that are responsible for hatching of embryos in fishes²⁷. Of the six duplicated genes in seahorse, five are highly expressed in the male brood pouch, suggesting that they may be involved in male pregnancy, possibly through rewiring of their regulatory network. The loss of pelvic fins in seahorse is associated with the evolution of an armour-like covering of its body and gain of an

elongated, flexible, substrate-gripping tail. By combining comparative genomics and gene-knockout experiments in zebrafish, we suggest that loss of *tbx4* may have a role in this phenotype in seahorse. The loss of mineralized teeth in seahorse is associated with the fusion of the jaws into a tube-like snout and a small mouth, which is extremely efficient in sucking small food items that are abundant in the benthic environment. In teleosts, P/Q-rich SCPP genes are involved in the mineralization of enameloid, which is the equivalent of enamel in tetrapods¹⁰. The seahorse genome does not contain any intact P/Q-rich SCPP genes that code for enamel matrix proteins, suggesting that the loss of these genes could have played a part in the loss of its mineralized teeth. Our analyses of the *H. comes* genome sequence and comparative genomics with other teleosts highlighted several genetic changes that may be involved in the evolution of the unique morphology of seahorses.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 March; accepted 2 November 2016.

- Leysen, H. *et al.* Musculoskeletal structure of the feeding system and implications of snout elongation in *Hippocampus reidi* and *Dunckerocampus dactylophorus*. *J. Fish Biol.* **78**, 1799–1823 (2011).
- Stölting, K. N. & Wilson, A. B. Male pregnancy in seahorses and pipefish: beyond the mammalian model. *BioEssays* **29**, 884–896 (2007).
- Wilson, A. B., Vincent, A., Ahnesjö, I. & Meyer, A. Male pregnancy in seahorses and pipefishes (family Syngnathidae): rapid diversification of paternal brood pouch morphology inferred from a molecular phylogeny. *J. Hered.* **92**, 159–166 (2001).
- Teske, P. R., Cherry, M. I. & Matthee, C. A. The evolutionary history of seahorses (Syngnathidae: *Hippocampus*): molecular data suggest a West Pacific origin and two invasions of the Atlantic Ocean. *Mol. Phylogenet. Evol.* **30**, 273–286 (2004).
- Near, T. J. *et al.* Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. *Proc. Natl Acad. Sci. USA* **110**, 12738–12743 (2013).
- Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
- Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, 2000).
- Bailey, X. *et al.* The loss of the hemoglobin H2S-binding function in annelids from sulfide-free habitats reveals molecular adaptation driven by Darwinian positive selection. *Proc. Natl Acad. Sci. USA* **100**, 5885–5890 (2003).
- MacArthur, D. G. *et al.* Loss of *ACTN3* gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nature Genet.* **39**, 1261–1265 (2007).
- Kawasaki, K. The SCPP gene family and the complexity of hard tissues in vertebrates. *Cells Tissues Organs* **194**, 108–112 (2011).
- Louchart, A. & Viriot, L. From snout to beak: the loss of teeth in birds. *Trends Ecol. Evol.* **26**, 663–673 (2011).
- Meredith, R. W., Zhang, G., Gilbert, M. T., Jarvis, E. D. & Springer, M. S. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science* **346**, 1254390 (2014).
- Deméré, T. A., McGowen, M. R., Berta, A. & Gatesy, J. Morphological and molecular evidence for a stepwise evolutionary transition from teeth to baleen in mysticete whales. *Syst. Biol.* **57**, 15–37 (2008).
- Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
- Yamanoue, Y., Setiamarga, D. H. & Matsuura, K. Pelvic fins in teleosts: structure, function and evolution. *J. Fish Biol.* **77**, 1173–1208 (2010).
- Kuiter, R. H. *Seahorses and their Relatives* (Aquatic Photographics, 2009).
- Harris, J. E. The role of the fins in the equilibrium of the swimming fish. II. The role of the pelvic fins. *J. Exp. Biol.* **15**, 32–47 (1938).
- Gosline, W. A. The evolution of some structural systems with reference to the interrelationships of modern lower teleostean fish groups. *Jpn. J. Ichthyol.* **27**, 1–28 (1980).
- Standen, E. M. Pelvic fin locomotor function in fishes: three-dimensional kinematics in rainbow trout (*Oncorhynchus mykiss*). *J. Exp. Biol.* **211**, 2931–2942 (2008).
- Tanaka, M. *et al.* Developmental genetic basis for the evolution of pelvic fin loss in the pufferfish *Takifugu rubripes*. *Dev. Biol.* **281**, 227–239 (2005).
- Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* **327**, 302–305 (2010).
- Naiche, L. A. & Papaioannou, V. E. Loss of *Tbx4* blocks hindlimb development and affects vascularization and fusion of the allantois. *Development* **130**, 2681–2693 (2003).
- Rodriguez-Esteban, C. *et al.* The T-box genes *Tbx4* and *Tbx5* regulate limb outgrowth and identity. *Nature* **398**, 814–818 (1999).
- Tamura, K., Yonei-Tamura, S. & Izpisua Belmonte, J. C. Differential expression of *Tbx4* and *Tbx5* in zebrafish fin buds. *Mech. Dev.* **87**, 181–184 (1999).
- Arora, R., Metzger, R. J. & Papaioannou, V. E. Multiple roles and interactions of *Tbx4* and *Tbx5* in development of the respiratory system. *PLoS Genet.* **8**, e1002866 (2012).
- Don, E. K. *et al.* Genetic basis of hindlimb loss in a naturally occurring vertebrate model. *Biol. Open* **5**, 359–366 (2016).
- Kawaguchi, M. *et al.* Evolution of teleostean hatching enzyme genes and their paralogous genes. *Dev. Genes Evol.* **216**, 769–784 (2006).
- Harlin-Cognato, A., Hoffman, E. A. & Jones, A. G. Gene cooption without duplication during the evolution of a male-pregnancy gene in pipefish. *Proc. Natl Acad. Sci. USA* **103**, 19407–19412 (2006).
- Whittington, C. M., Griffith, O. W., Qi, W., Thompson, M. B. & Wilson, A. B. Seahorse brood pouch transcriptome reveals common genes associated with vertebrate pregnancy. *Mol. Biol. Evol.* **32**, 3114–3131 (2015).
- Kawaguchi, M., Tomita, K., Sano, K. & Kaneko, T. Molecular events in adaptive evolution of the hatching strategy of ovoviparous fishes. *J. Exp. Zool. B Mol. Dev. Evol.* **324**, 41–50 (2015).
- Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
- Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
- Venkatesh, B. *et al.* Ancient noncoding elements conserved in the human genome. *Science* **314**, 1892 (2006).
- Navratilova, N. *et al.* Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.* **327**, 526–540 (2009).
- Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature Genet.* **40**, 158–160 (2008).
- Attanasio, C. *et al.* Fine tuning of craniofacial morphology by distant-acting enhancers. *Science* **342**, 1241006 (2013).
- McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011).
- Sabherwal, N. *et al.* Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. *Hum. Mol. Genet.* **16**, 210–222 (2007).
- Shears, D. J. *et al.* Mutation and deletion of the pseudoautosomal gene *SHOX* cause Leri-Weill dyschondrosteosis. *Nature Genet.* **19**, 70–73 (1998).
- Indjeian, V. B. *et al.* Evolving new skeletal traits by cis-regulatory changes in bone morphogenetic proteins. *Cell* **164**, 45–56 (2016).
- Lin, Q., Lin, J. & Zhang, D. Breeding and juvenile culture of the lined seahorse, *Hippocampus erectus* Perry, 1810. *Aquaculture* **277**, 287–292 (2008).


Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the Laboratory Animal Center of Sun Yat-Sen University for providing experimental facilities and the Agency for Science, Technology and Research (A*STAR) Computational Resource Centre for use of its high-performance computing facilities. This research was supported by the National Science Fund for Excellent Young Scholars (41322038), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA13020103), the Youth Foundation of National High Technology Research and Development Program (863 Program) (2015AA020909), the Outstanding Youth Foundation in Guangdong Province (S2013050014802), the National Natural Science Foundation of China (41576145), the National Key Basic Research Program of China (2015CB452904), the Special Project on the Integration of Industry, Education and Research of Guangdong Province (no. 2013B090800017) and the Biomedical Research Council of A*STAR, Singapore.

Author Contributions Q.L., A.M. and B.V. designed the scientific objectives. Q.L., B.V., A.M., Q.S. and Y.Z. oversaw the project. H.Z., G.Q., B.V. and Y.Z. collected samples for sequencing DNA and RNA. M.X., Y.Y., J.M. and Q.G. performed genome sequencing, assembly and annotation. S.F., J.M. and Y.Y. performed phylogenomic analysis and molecular evolutionary rate analysis. S.F. and M.X. characterized repetitive sequences and GC content. S.F., R.F.S., P.X., V.R. and B.V. annotated and analysed Hox clusters. V.R. and B.V. annotated and analysed SCPP genes. A.P.L., V.R., Z.W.L. and B.V. performed CNE analysis and functional assay of zebrafish CNEs. Y.Z., M.X., C.Z. and D.S. assembled and annotated RNA-seq data. H.M.G. and S.F. interpreted RNA-seq results and designed the qRT-PCR experiment. Y.Z., H.Z. and X.W. performed qRT-PCR to validate the expression levels of transcripts. Y.Z., M.X., H.Z. and V.R. analysed the *patristacin* gene family. Y.Z., Q.L., J.M.W., R.F.S. and A.M. performed *tbx4* knockout analysis. Y.Q., J.B., C.B., Y.S. and X.Z. were involved in data analysis. L.H., G.L., W.L., Z.G., K.W. and H.Q. participated in the discussions related to data analysis. Q.L., Y.Z., S.F., H.M.G., A.M. and B.V. wrote the manuscript with input from all other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Q.S. (shiqiong@genomics.cn), A.M. (axel.meyer@uni-konstanz.de) or B.V. (mcbbv@imcb.a-star.edu.sg).

Reviewer Information Nature thanks C. Amemiya, S. Burgess, K. Worley and the other anonymous reviewer(s) for their contribution to the peer review of this work.

 This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

Genome sequencing and assembly. Genomic DNA of a single male *H. comes* was used to construct eleven libraries including short-insert (170 bp, 500 bp, 800 bp) and mate-paired (2 kb, 5 kb, 10 kb, 20 kb) libraries and sequenced on the Illumina HiSeq 2000 sequencing platform. In total, we obtained around 218 Gb of raw sequence data (Supplementary Table 1.1). The genome was assembled using SOAPdenovo2.04 (ref. 42) with default parameters. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

RNA sequencing and analysis. In total, 19 RNA-seq libraries were constructed, including two libraries from combined soft tissues (brain, gills, intestine, liver and muscle) from a male and a female *H. comes* (Supplementary Table 2.1); and 17 libraries of five developmental stages of embryos and different stages of brood pouch development such as the juvenile stage, rudimentary stage, pre-pregnancy stage, pregnancy stage, and post pregnancy stage, using RNA from the lined seahorse (*Hippocampus erectus*) (Supplementary Information, section 2). All libraries were prepared using Illumina TruSeq RNA sample preparation kit according to the manufacturer's instructions (Illumina, San Diego, CA, USA) and sequenced using Illumina HiSeq 2000 platform. The RNA-seq reads were either *de novo* assembled using Trinity⁴³ or mapped to the *H. comes* genome using TopHat⁴⁴ with default parameters, and subsequently analysed using in-house Perl scripts. The differential expression of genes at different stages of brood pouch development was determined using the method developed previously⁴⁵. The RNA-seq results were validated using qRT-PCR, with five biological replicates for each stage. All data were expressed as mean \pm standard error of mean and were evaluated by one-way ANOVA followed by Tukey's honestly significant difference test for adjusting *P* values from multiple comparisons. Results were considered to be statistically significant for *P* values < 0.05 .

Genome annotation. Annotation of the *H. comes* genome was carried out using the Ensembl gene annotation pipeline which integrated *ab initio* gene predictions and evidence-based gene models. Briefly, protein sequences of *D. rerio*, *G. aculeatus*, *O. latipes*, *T. rubripes* and *T. nigroviridis* were downloaded from Ensembl (release 75) and mapped to the genome using TblastN⁴⁶ with the parameter “-evalue 1E-5”. Second, high scoring segment pairs (HSPs) from blast were concatenated using Solar (in-house software, version 0.9.6). Third, the concatenated segments were aligned using GeneWise⁴⁷ to refine the gene models. Finally, we filtered the alignments that showed alignment rates less than 50% of the full-length copies and filtered redundant alignments based on the GeneWise score. In addition, *H. comes* transcripts (female_transcript and male_transcript) and *H. erectus* transcripts (Juv_brain, Juv_body, Rud_testis and PreP_pouch) were used to assist in the gene model prediction. We annotated the predicted gene models using Swiss-Prot, TrEMBL, NCBI NR database, and KEGG databases (Supplementary Table 3.4).

Expansion and contraction of gene families. We used CAFE (version 2.1), a program for analysing gene family expansion and contraction under maximum likelihood framework. The gene family results from TreeFam pipeline and the estimated divergence time between species were used as inputs. We used the parameters “-p 0.01, -r 10000, -s” to search the birth and death parameter (λ) of genes, calculated the probability of each gene family with observed sizes using 10,000 Monte Carlo random samplings, and reported birth and death parameters in gene families with probabilities less than 0.01. For the gene family expansion and contraction analysis in *H. comes*, we first filtered out gene families without homology in the SWISS-PROT database to reduce the potential false positive expansions or contractions caused by gene prediction. The families that contained sequences that have multiple functional annotations were also removed (Supplementary Tables 4.1 and 4.2).

Phylogenetic analysis. We obtained 4,122 one-to-one orthologous genes from the gene family analysis (Supplementary Information, section 4.1). The protein sequences of one-to-one orthologous genes were aligned using MUSCLE⁴⁸ with the default parameters. We then filtered the saturated sites and poorly aligned regions using trimAl (ref. 49) with the parameters “-gt 0.8 -st 0.001 -cons 60”. After trimming the saturated sites and poorly aligned regions in the concatenated alignment, 2,128,000 amino acids were used for the phylogenomic analysis. The trimmed protein alignments were used as a guide to align corresponding coding sequences (CDSs). The aligned protein and the fourfold degenerate sites in the CDSs were each concatenated into a super gene using an in-house Perl script.

The phylogenomic tree was reconstructed using RAxML version 8.1.19 (ref. 50) based on concatenated protein sequences. Specifically, we used the PROTGAUTO parameter to select the optimal amino acid substitution model, specified spotted gar as the outgroup, and evaluated the robustness of the result using 100 bootstraps. To compare the neutral mutation rate of different species, we also generated a phylogeny based on fourfold degenerate sites. The phylogenomic topology was used as input and the “-f e” option in RAxML was used to optimize the branch lengths of the input tree using the alignment of fourfold degenerate sites under the general time reversible (GTR) model as suggested by ModelGenerator

version 0.85 (ref. 51). We calculated the pairwise distances to the outgroup (spotted gar) based on the optimized branch length of the neutral tree using the cophenetic.phylo module in the R-package APE⁵². The Bayesian relaxed-molecular clock (BRMC) method, implemented in the MCMCTree program⁵³, was used to estimate the divergence time between different species. The concatenated CDS of one-to-one orthologous genes and the phylogenomics topology were used as inputs. Two calibration time points based on fossil records, *O. latipes*–*T. nigroviridis* (~96.9–150.9 million years ago (Mya)), and *D. rerio*–*G. aculeatus* (~149.85–165.2 Mya) (<http://www.fossilrecord.net/dateacade/index.html>), were used as constraints in the MCMCTree estimation. Specifically, we used the correlated molecular clock and REV substitution model in our calculation. The MCMC process was run for 5,000,000 steps and sampled every 5,000 steps. MCMCTree suggested that *H. comes* diverged from the common ancestor of stickleback, Nile tilapia, platyfish, fugu, and medaka approximately 103.8 Mya, which corresponds to the Cretaceous period.

Analysis of OR genes. We downloaded protein sequences of 1,417 OR gene family members from NCBI and mapped them to *H. comes* genome using Tblastn with “E-value $\leq 1e-10$ ” and “alignment rate ≥ 0.5 ”. Solar (in-house software, version 0.9.6) was used to join high-scoring segment pairs (HSPs) between each pair of protein mapping results. We retained alignments with an alignment rate of more than 70% and a mapping identity of more than 40%. Subsequently, the protein sequences were mapped to the genome using GeneWise and extended 280 bp upstream and downstream to define integrated gene models. For phylogenetic analysis, protein sequences were aligned using MUSCLE and a JTT+gamma model was used in a maximum-likelihood analysis using PhyML to construct a phylogenetic tree.

Evidence for loss of *tbx4* in *H. comes*. The synteny analysis of *tbx2b*–*tbx4*–*brip1* region of *H. comes*, stickleback, fugu and zebrafish using Vista shows that *tbx4* was lost in *H. comes* (Fig. 3). To exclude the scenario that the absence of *tbx4* in the *H. comes* genome sequence is due to an assembly error, we first validated the micro-synteny region of *tbx2b*–*tbx4*–*brip1* region in *H. comes* using a PCR-based genomic walk strategy. Briefly, 28 primer pairs (Supplementary Table 9.1) were designed for overlapping amplicons to ‘walk’ from the end of *tbx2b* to the start of *brip1*. Amplicon size and partial end sequencing of these products did not indicate any anomalies in the assembly of the *H. comes* *tbx4* ‘ghost’ locus.

In addition, we carried out the following analyses: (1) searched the *H. comes* genome (TblastN) using Tbx4 protein from zebrafish and Nile tilapia and were unable to find a *tbx4* gene; (2) searched the *H. comes* genome using only the domain sequence of Tbx4 protein but were unable to find a *tbx4* gene; (3) searched *H. comes* and *H. erectus* transcriptome data for *tbx4* (TblastN) using Tbx4 protein from zebrafish and Nile tilapia but were unable to find any matching transcript; (4) searched *H. comes* and *H. erectus* transcriptome data with the domain sequence as well and did not find any remnant of a *tbx4* gene; and (5) predicted CNEs in the ‘ghost’ *tbx4* locus of *H. comes* using the fugu *tbx4* locus as the reference (base) (Supplementary Fig. 9.3). We used the CNEs present in the other fish genome loci (that were absent in *H. comes*) to search the *H. comes* genome to rule out the possibility that they may be present elsewhere in the genome. We were unable to find any of these CNEs in the *H. comes* genome. Finally, we conducted degenerate PCR experiments to ascertain if the *tbx4* gene is missing in *H. comes*. Using a combination of four forward and two reverse primers (Supplementary Table 9.1), we checked for the presence of *tbx4* in seven species of *Hippocampus* (including *H. comes* and *H. erectus*), five species of pipefish (four from the genus *Syngnathus* and one species of *Corythoichthys*) (all from the family Syngnathidae that lack pelvic fins); ghost pipefish (*Solenostomus*) and the trumpetfish (*Aulostomidae*) which are closely related to the Syngnathidae but possess pelvic fins; and five other teleost species that possess pelvic fins (Supplementary Figs 9.1 and 9.2).

Generation of mutant *tbx4* zebrafish. We used a CRISPR–Cas9 strategy to generate a *tbx4* mutant zebrafish line. Two guide RNAs (gRNAs) were designed targeting zebrafish *tbx4* in the 5' end of the sequence that is upstream of or within the DNA-binding TBOX domain (Supplementary Fig. 9.4). gRNAs were cloned using synthesized oligonucleotides into the pT7gRNA vector as described previously⁵⁴ (oligonucleotide sequences given in Supplementary Table 9.2). gRNAs were synthesized from this vector after linearization with BamHI–HF (NEB R3136T), transcribed using the MEGascript T7 Transcription Kit (Thermo Fischer Scientific AM1334) and purified using the mirVana miRNA isolation kit (Thermo Fischer Scientific AM1560). Cas9 mRNA was synthesized from the Cs2+Cas9 vector using the mMessage mMachine Sp6 Transcription Kit (Thermo Fischer Scientific AM1340) and purified using the RNA cleanup protocol from the RNeasy mini kit (Qiagen 74104).

Zebrafish from a wild caught strain were injected at the one-cell stage with ~50 ng gRNA and ~90 ng Cas9 RNA. These F0 fish were raised to maturity and genotyped using fin clipping, DNA isolation and PCR spanning the target site (genotyping primers given in Supplementary Table 9.2). PCR products were analysed for mutations as described previously⁵⁴ using T7 endonuclease (NEB M0302L). Mosaic mutant F0 fish were outcrossed to AB wild-type fish and embryos were batch genotyped for transmission of the mutation using PCR and T7 endonuclease. Mutant PCR products were cloned into the pGEM-T vector

(Promega, Madison, WI) and sequenced to identify carrier fish transmitting a frameshift mutation. These carrier fish were crossed again to AB wild type and the resulting F1 fish were raised to maturity. The F1 were genotyped using fin clipping, DNA isolation, PCR, T7 endonuclease to identify heterozygous mutant fish followed by cloning and sequencing of the mutant PCR products to validate presence of the frameshift allele. The CRISPR–Cas9 mutation strategy is schematically shown in Extended Data Fig. 5.

In the F0 mutant *tbx4* fish we observed pelvic fin loss at low frequency. gRNA#1 gave 3/42 fish with either double- or single-sided pelvic fin loss whereas 1/34 had single-sided pelvic fin loss for gRNA#2 (Extended Data Fig. 5). We observed mutant allele transmission for both gRNA#1 and gRNA#2 but failed to identify a deletion leading to a frameshift mutation for gRNA#2 so no stable line was generated for this CRISPR. For gRNA#1 we identified several frameshift mutants, one of which was further analysed. This mutant has a deletion/replacement mutation in which eight nucleotides are replaced by three nucleotides, leading to an effective 5 bp deletion and the introduction of a frameshift mutation (Extended Data Fig. 5). This mutation introduces a downstream STOP codon leading to a severely truncated protein lacking the DNA binding domain (Supplementary Figs 9.4 and 9.5). The mutant line is maintained on an AB wild-type background. **Loss of CNEs.** Using zebrafish as the reference genome, whole-genome alignments of six teleost fishes were generated. The soft-masked genome sequence for zebrafish (Zv9, April 2010) was downloaded from the Ensembl release-75 FTP site. The following soft-masked genome sequences were downloaded from the UCSC Genome Browser: stickleback (gasAcu1, February 2006), fugu (fr3, October 2011), medaka (oryLat2, October 2005), Nile tilapia (oreNil2, February 2012). The *H. comes* genome sequence (hipCom0) was repeat-masked using WindowMasker (from NCBI BLAST+ package v.2.2.28) with additional parameter “-dust true”. About 32% (158.1/501.6 Mb) of the *H. comes* genome was masked using this method.

Only chromosome sequences of zebrafish were aligned while unplaced scaffolds were excluded. The reference (zebrafish) genome was split into 21 Mb sequences with 10-kb overlap, while the percomorph fish genomes (*H. comes*, stickleback, fugu, medaka and Nile tilapia) were split into 10 Mb sequences with no overlap. Pairwise alignments were carried out using Lastz v.1.03.54 (ref. 55) with the following parameters: –strand=both–seed=12of19–notransition–chain–gapped–gap=400,30–hspthresh=3000–gappedthresh=3000–inner=2000–masking=50–ydrop=9400–scores=HoxD55.q–format=txt. Coordinates of split sequences were restored to genome coordinates using an in-house Perl script. The alignments were reduced to single coverage with respect to the reference genome using UCSC Genome Browser tools ‘actChain’ and ‘chainNet’. Multiple alignments were generated using Multiz.v11.2/roast.v3 (ref. 56) with the tree topology “(Zv9 (hipCom0 ((fr3 gasAcu1) (oryLat2 oreNil2)))”.

Fourfold degenerate (4D) sites of zebrafish genes (Ensembl release-75) were extracted from the multiple alignments. These 4D sites were used to build a neutral model using PhyloFit in the rphast v.1.5 package⁵⁷ (general reversible “REV” substitution model). PhastCons was then run in rho-estimation mode on each of the zebrafish chromosomal alignments to obtain a conserved model for each chromosome. These conserved models were averaged into one model using PhyloBoot. Subsequently, conserved elements were predicted in the multiple alignments using PhastCons with the following inputs and parameters: the neutral and conserved models, target coverage of input alignments = 0.3 and average length of conserved sequence = 45 bp. To assess the sensitivity of this approach in identifying functional elements, the PhastCons elements were compared against zebrafish protein-coding genes. Eighty per cent of protein-coding exons (197,508/245,556 exons) were overlapped by a conserved element (minimum coverage 10%), indicating that the identification method was fairly sensitive.

A CNE was considered present in a percomorph genome if it showed coverage of at least 30% with a zebrafish CNE in Multiz alignment. To identify CNEs that could have been missed in the Multiz alignments due to rearrangements in the genomes, or due to partitioning of the CNEs among teleost fish duplicate genes, we searched the zebrafish CNEs against the genome of the percomorph using BLASTN ($E < 1 \times 10^{-10}$; $\geq 80\%$ identity; $\geq 30\%$ coverage). Those CNEs that had no significant match in a percomorph genome were considered as missing in that genome. To account for CNEs that might have been missed due to sequencing gaps, we identified gap-free syntenic intervals in zebrafish and the percomorph genomes, and generated a set of CNEs that were missing from these intervals. These CNEs represent a high-confidence set of CNEs missing in the percomorph fishes and thus were used for further analysis. Functional enrichment of genes associated with CNEs was carried out using the GREAT software⁵⁸ with each CNE assigned to the genes with the nearest transcription start site and within 1 Mb in the zebrafish genome, and significantly enriched functional categories identified based on a hypergeometric test of genomic regions (false discovery rate (FDR) q value < 0.05). We identified the statistically significant gene ontology biological process terms, molecular function terms and zebrafish phenotype descriptions of the genes that are associated with CNEs.

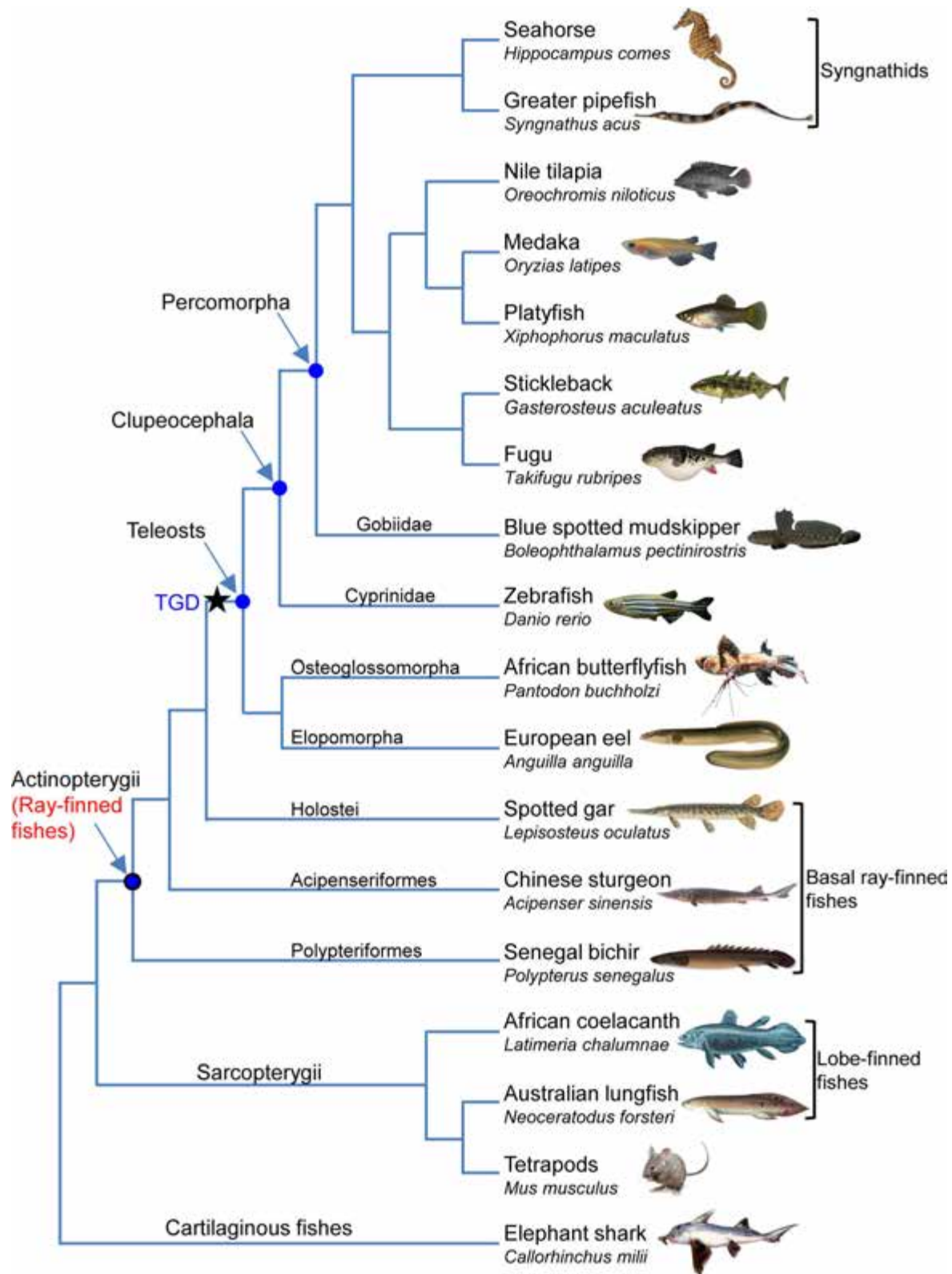
We also predicted CNEs in the Hox clusters of *H. comes* and other representative teleost fishes using the global alignment program MLAGAN. Orthologous Hox clusters were aligned using MLAGAN with zebrafish as the reference sequence and CNEs were predicted using VISTA.

Functional assay of CNEs. Seven representative zebrafish CNEs that have been lost in *H. comes* (the largest among the lost CNEs) were assayed for enhancer activity in transgenic zebrafish using GFP as the reporter gene. The CNEs were amplified by PCR using zebrafish genomic DNA as template. The products were cloned into a miniTol2 transposon donor plasmid linked to the mouse *cFos* (McFos) basal promoter and the coding sequence of GFP. Transposase mRNA was generated by transcribing cDNA *in vitro* using the mMESSAGE mMACHINE T7 kit (Ambion; Life Technologies). The CNE-containing McFos-miniTol2 construct and transposase mRNA were co-injected into the yolk of zebrafish embryos at the one to two-cell stage. Each CNE construct was injected into 250–350 embryos and the injections were repeated on two days. The embryos were reared at 28 °C, and GFP was observed at 24, 48 and 72 h post-fertilization (hpf). The survival rate of the embryos post-injection was 70–80%. Consistent GFP expression in at least 20% of F0 embryos was considered as specific expression driven by a CNE. Such embryos were reared to maturity and mated with wild type zebrafish to produce F1 lines. The expression of GFP in F1 embryos was observed under a compound microscope fitted for epifluorescence (Axio imager M2; Carl Zeiss, Germany) and photographed using an attached digital microscope camera (Axiocam; Carl Zeiss, Germany). Pigmentation was inhibited by maintaining zebrafish embryos in 0.003% N-phenylthiourea (Sigma-Aldrich, Sweden) from 8 hpf onwards. Consistent GFP expression observed in at least three lines of F1 fishes was considered as the specific expression driven by a CNE.

All animals were cared for in strict accordance with National Institutes of Health (USA) guidelines. The zebrafish gene knockout protocol was approved by the Institutional Animal Care and Use Committee of Sun Yat-Sen University. The zebrafish transgenic assay protocol was approved by the Institutional Animal Care and Use Committee of Biological Resource Centre, A*STAR, Singapore.

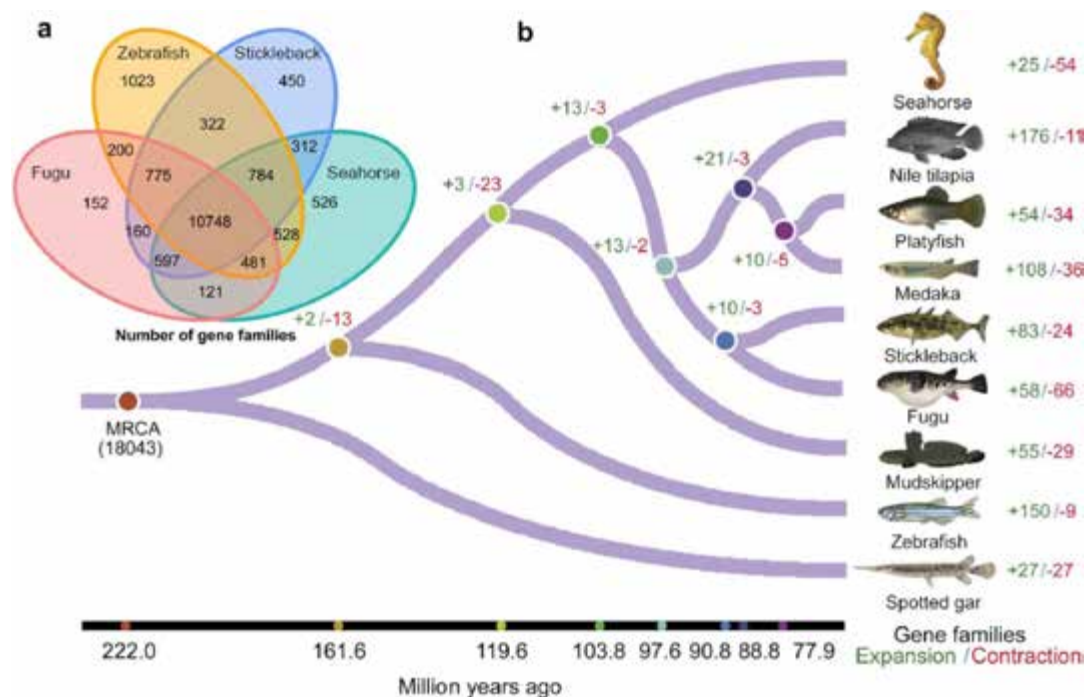
Data availability statement. The tiger tail seahorse (*H. comes*) whole-genome sequence has been deposited in the DDBJ/EMBL/GenBank database under accession number LVHJ00000000. RNA-seq reads for *H. erectus* and *H. comes* have been deposited in the NCBI Sequence Read Archive under accession numbers SRA392578 and SRA392580, respectively.

42. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
43. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnol.* **29**, 644–652 (2011).
44. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
45. Yu, X., Lin, J., Zack, D. J. & Qian, J. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* **34**, 4925–4936 (2006).
46. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
47. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
48. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
49. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
50. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
51. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).
52. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
53. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
54. Jao, L. E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl Acad. Sci. USA* **110**, 13904–13909 (2013).
55. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State Univ. (2007).
56. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
57. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).
58. McLean, C. Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotechnol.* **28**, 495–501 (2010).
59. Bian, C. *et al.* The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Sci. Rep.* **6**, 24501 (2016).
60. Kawasaki, K. & Amemiya, C. T. SCPP genes in the coelacanth: tissue mineralization genes shared by sarcopterygians. *J. Exp. Zool. B Mol. Dev. Evol.* **322**, 390–402 (2014).
61. Venkatesh, B. *et al.* Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179 (2014).



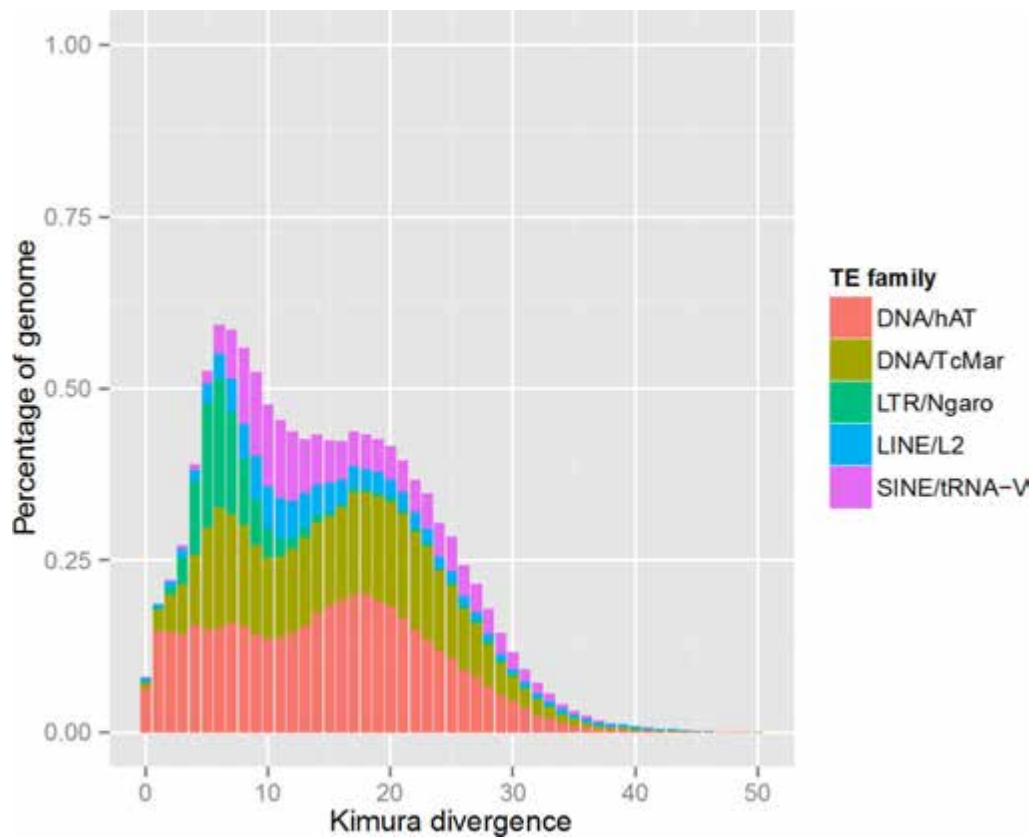
Extended Data Figure 1 | Phylogenetic relationships of ray-finned fishes discussed in this study. Phylogenetic relationships of ray-finned fishes depicted here are based on the current study and ref. 59. Ray-finned fishes (Actinopterygii) are divided into basal ray-finned fishes (Polypteriformes, Acipenseriformes and Holostei) and teleosts.

The latter comprise ~99% of the extant ray-finned fishes. The star represents the teleost-specific genome duplication (TGD) event that occurred in the common ancestor of all teleost fishes. Syngnathids (seahorse and pipefish) display the unique phenomenon of ‘male pregnancy’.

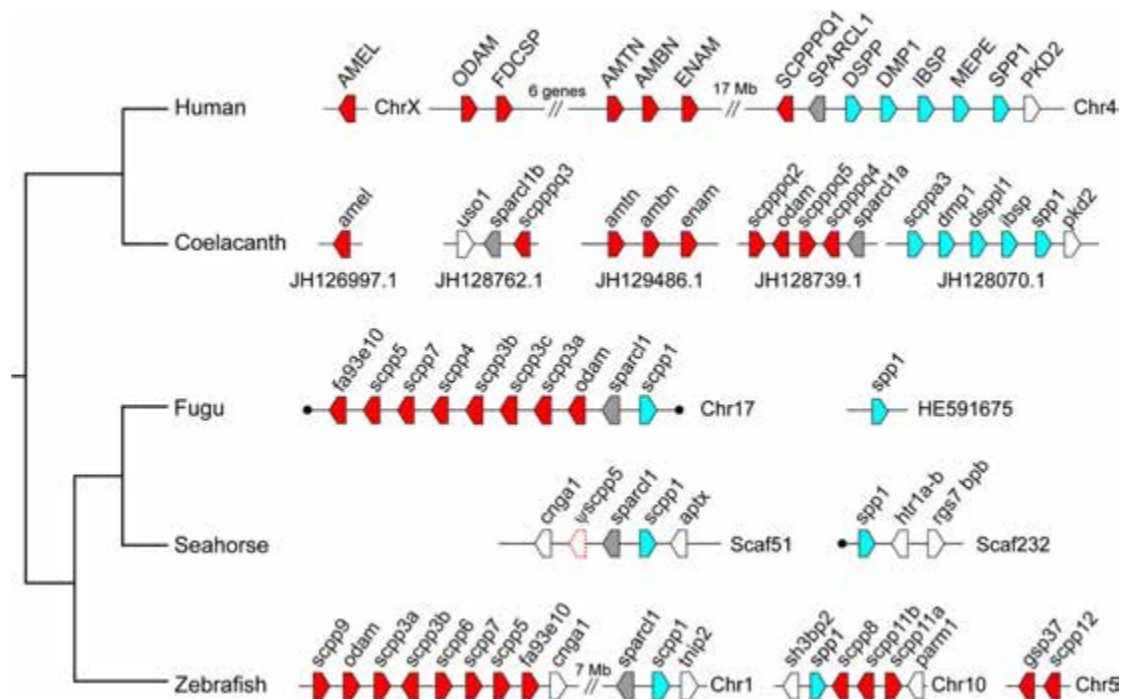


Extended Data Figure 2 | Number of gene families in various teleosts and the spotted gar. a, Venn diagram of shared orthologous gene families in seahorse (*H. comes*), fugu, zebrafish and stickleback. **b,** The phylogeny and divergence times of seahorse and other teleost fishes based on analysis

of genome-wide one-to-one orthologous protein sequences. The numbers at nodes indicate the number of gene families expanded and contracted at different evolutionary time points.

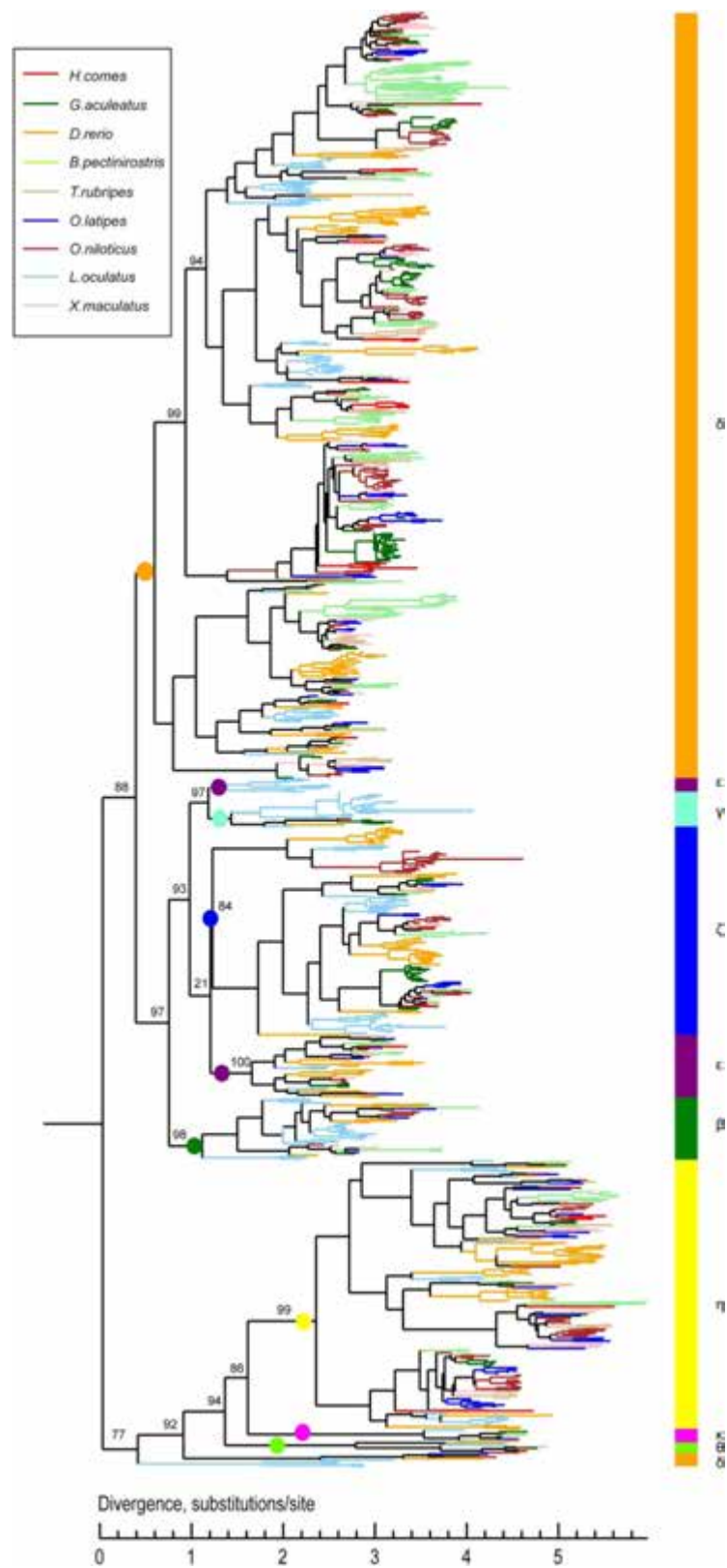


Extended Data Figure 3 | Divergence distribution of transposable elements compared to consensus in the transposable element library. The divergence rate was calculated between the identified transposable elements (TEs) in the *H. comes* genome and the consensus sequence in the transposable element library.

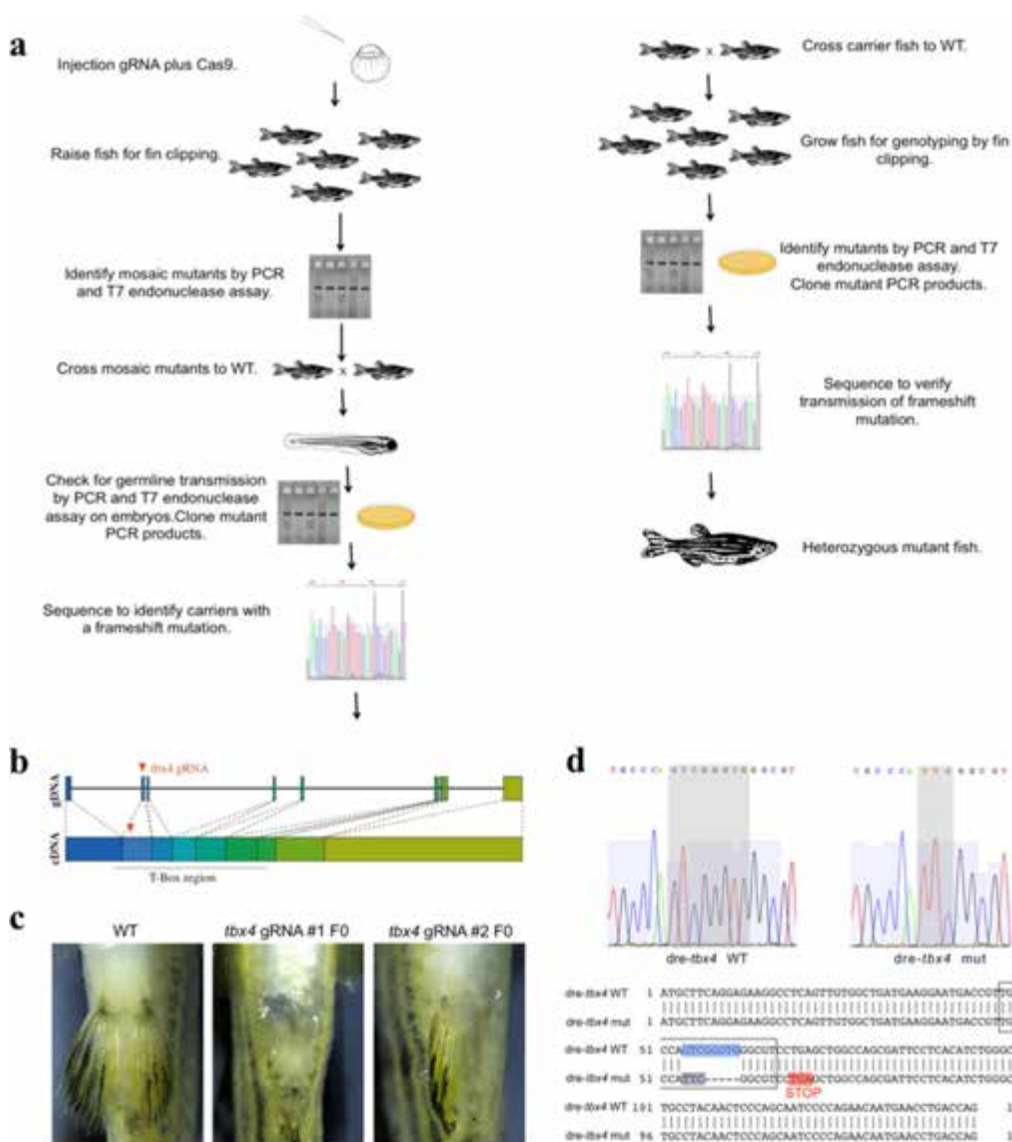


Extended Data Figure 4 | SSCP genes in *H. comes* and other jawed vertebrates. Gene loci for human, coelacanth and zebrafish were adapted from other publications^{60,61}. *sparr1*, which is the ancestral gene that gave rise to SSCP genes is shown in grey; P/Q-rich SSCP genes are shown

in red; acidic SSCP genes are shown in blue. In seahorse, *scpp5* is a pseudogene and is denoted by ψ . Owing to space constraints, the P/Q-rich SSCP genes encoding milk casein and salivary proteins in human have been omitted. Black circles mark the ends of scaffolds.

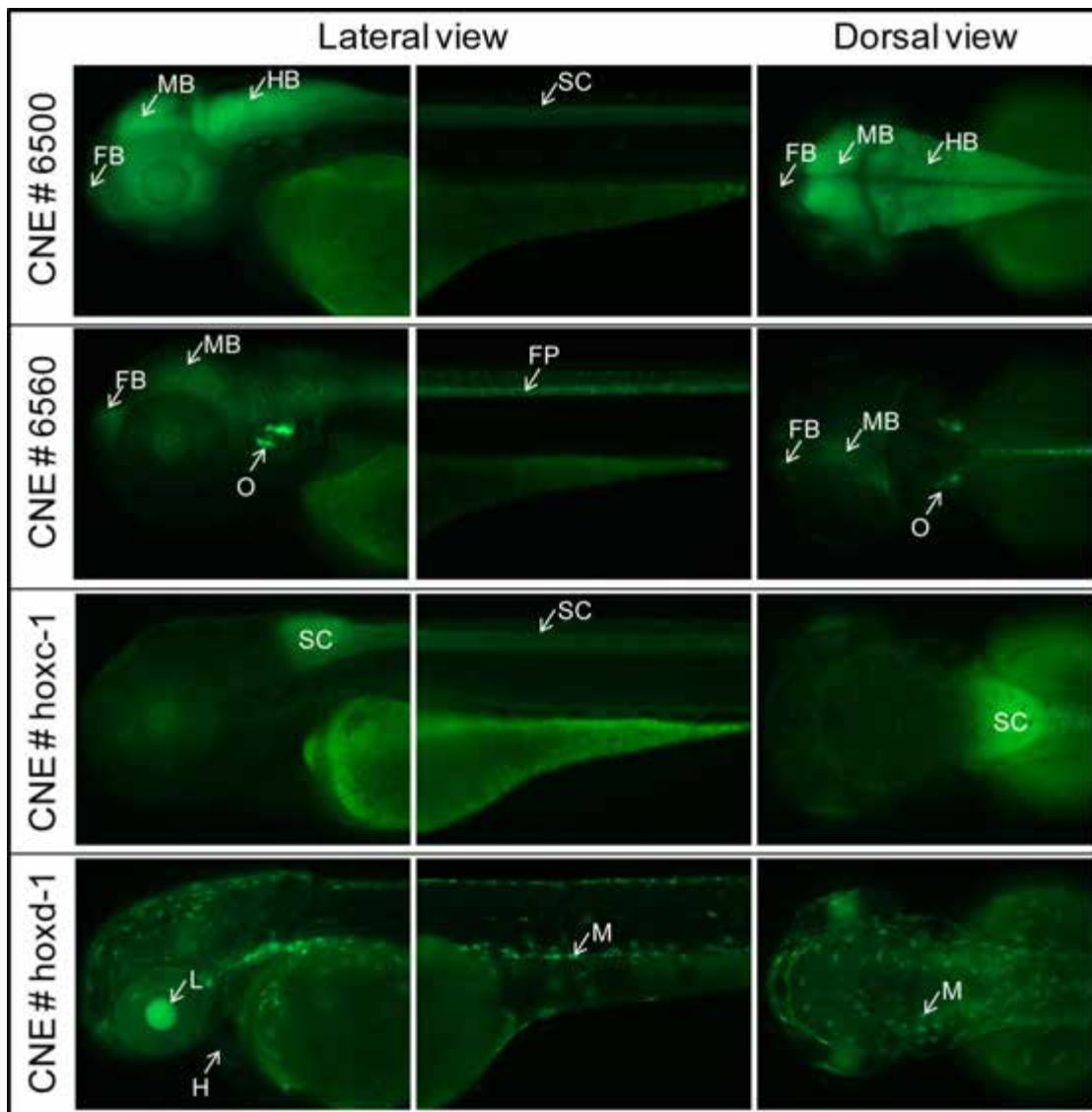


Extended Data Figure 5 | Maximum-likelihood phylogenetic tree of OR genes in *H. comes* and other ray-finned fishes.



Extended Data Figure 6 | CRISPR-Cas9 mediated knockdown of *tbx4* in zebrafish. **a**, CRISPR-Cas9 mutagenesis strategy. **b**, CRISPR-Cas9 sites targeted in zebrafish *tbx4* gene. **c**, Loss of function *tbx4* phenotypes in F0 mosaic mutants. Pelvic fin loss was observed with low frequency in F0 mosaic mutant fish. Frequency of animals with either single- or double-sided loss of pelvic fins was 3/42 for gRNA#1 and 1/34 for gRNA#2. **d**, Identification of zebrafish *tbx4* mutant line. Top shows

sequencing chromatograms of wild-type (left) and mutant (right) alleles. Bottom shows alignment of *tbx4* exon 2 from wild-type and mutant. The region for which the chromatograms are shown is indicated with a box. In the mutant a deletion (indicated in blue in the wild-type sequence)/substitution (indicated in lilac in the mutant sequence) was identified. The deletion/substitution area is indicated with a grey box in the chromatograms.



Extended Data Figure 7 | Reporter gene expression pattern driven by zebrafish CNEs that are lost in *H. comes*. Lateral and dorsal views of 72 h post-fertilization F1 transgenic zebrafish embryos. The lost CNEs (#6500, #6560, #hoxc-1 and #hoxd-1) were assayed for their reporter gene

expression potential in transgenic zebrafish. FB, forebrain; FP, floor plate; H, heart; HB, hindbrain; L, lens; M, melanocytes; MB, midbrain; O, otic vesicle; SC, spinal cord.

Electric-field-stimulated protein mechanics

Doeke R. Hekstra^{1†}, K. Ian White¹, Michael A. Socolich¹, Robert W. Henning², Vukica Šrajer² & Rama Ranganathan^{1,3}

The internal mechanics of proteins—the coordinated motions of amino acids and the pattern of forces constraining these motions—connects protein structure to function. Here we describe a new method combining the application of strong electric field pulses to protein crystals with time-resolved X-ray crystallography to observe conformational changes in spatial and temporal detail. Using a human PDZ domain (LNX2^{PDZ2}) as a model system, we show that protein crystals tolerate electric field pulses strong enough to drive concerted motions on the sub-microsecond timescale. The induced motions are subtle, involve diverse physical mechanisms, and occur throughout the protein structure. The global pattern of electric-field-induced motions is consistent with both local and allosteric conformational changes naturally induced by ligand binding, including at conserved functional sites in the PDZ domain family. This work lays the foundation for comprehensive experimental study of the mechanical basis of protein function.

The fundamental biological properties of proteins—binding, catalysis and allosteric communication—emerge from a global pattern of interactions between all constituent atoms. Often, this pattern is organized in the tertiary structure so as to produce the concerted motions of amino acid residues, defining transitions between a small number of functional states. This conformational cycling within proteins and protein complexes draws analogies to macroscopic machines¹ and lies at the heart of many biological processes: DNA replication², metabolism^{3,4}, transport⁵, cellular motility⁶ and signal transduction⁷. Even without conformational changes, functional states of proteins can have a different pattern and extent of rigidity^{8,9}—entropic variations that also influence state transitions¹⁰. Thus, the biology of proteins is deeply connected to their mechanics: the motions a protein can perform and the forces constraining these motions. As in macroscopic machines¹, a comprehensive description of internal mechanics is the key to explaining how structure leads to function¹¹. Unlike conventional machines, however, proteins are marginally stable evolved materials whose mechanics are governed by weak, heterogeneously cooperative interactions for which we as yet have no good physical models.

Current biophysical methods provide an incomplete basis for making mechanical models of proteins. NMR spectroscopy¹² and room-temperature crystallography¹³ provide information on the structure and dynamics of local environments of atoms, and have been used to characterize weakly populated excited states of proteins^{3,13,14}. However, the complexity of disentangling conformational transitions occurring on multiple timescales, the difficulty of directly seeing collective motions, and the inability to generally relate the measured parameters to physical forces limit our understanding. Single molecule force spectroscopy can relate global conformational transitions to applied forces, but is limited in providing the atomic detail required to define the underlying intramolecular mechanics¹⁵. Time-resolved crystallography (TRX) offers, in principle, a direct route to observing concerted motions with high temporal and spatial resolution¹⁶. However, TRX traditionally relies on photoexcitation of bound chromophores to induce motions in proteins¹⁷. Such excitation is not generally applicable, acts at a fixed location, is not tunable, and deposits an amount of

energy that far exceeds the typical energetic changes involved in protein conformational transitions.

Here, we describe a new method for studying protein mechanics and its application to a model system—a PDZ domain—which shows both local and allosteric functional properties¹⁸. The method, electric field-stimulated X-ray crystallography (EF-X), combines the use of strong electric field pulses to drive motions within protein crystals with simultaneous readout by fast X-ray pulses. EF-X satisfies the key characteristics required for a general mechanical analysis of proteins: (1) the application of forces of controlled magnitude, direction and duration; (2) the existence of defined, well-distributed actuators (the charges) on which the forces act; and (3) readout of conformational changes with high spatial and temporal resolution. We show that EF-X can reveal protein motions associated with biological function and permits direct refinement of the atomic structures of low-lying excited states. This work initiates a path towards a full description of protein mechanics.

Theoretical and practical considerations

The idea of EF-X is simple; many elementary charges and local dipoles are present in proteins (Fig. 1a), and with the application of sufficiently large external electric fields, it should be possible to exert forces on them that cause motions of atoms throughout the protein structure. If the electric field can be applied in conjunction with timed X-ray diffraction in protein crystals, it should be possible to observe all of these motions in high spatial and temporal detail (Fig. 1b). To implement the idea, we began with a few design considerations. Theoretical calculations suggest that electric field strengths of $\sim 1,000,000 \text{ V cm}^{-1}$ are in the right range to drive subtle motions of atoms within proteins that can be observed through high-resolution diffraction methods (Methods and Extended Data Table 1). Fields of 1 MV cm^{-1} are dangerously large from a laboratory point of view, but are close to physiological; for example, 0.125 MV cm^{-1} corresponds to $\sim 100 \text{ mV}$ across a cell membrane. Such voltages influence conformational transitions in proteins such as ion channels¹⁹ and G-protein-coupled receptors²⁰, and are consistent with biological relevance. In general, the basic premise of EF-X is that features corresponding to the biologically relevant reaction coordinate(s) of proteins are enriched in the low-lying energetic states around

¹Green Center for Systems Biology, UT Southwestern Medical Center, 6001 Forest Park Road, Dallas, Texas 75390, USA. ²Center for Advanced Radiation Sources, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, USA. ³Departments of Biophysics and Pharmacology, UT Southwestern Medical Center, 6001 Forest Park Road, Dallas, Texas 75390, USA.

[†]Present address: Department of Molecular and Cellular Biology and School of Engineering and Applied Sciences, Harvard University, 52 Oxford Street, Cambridge, Massachusetts 02138, USA.

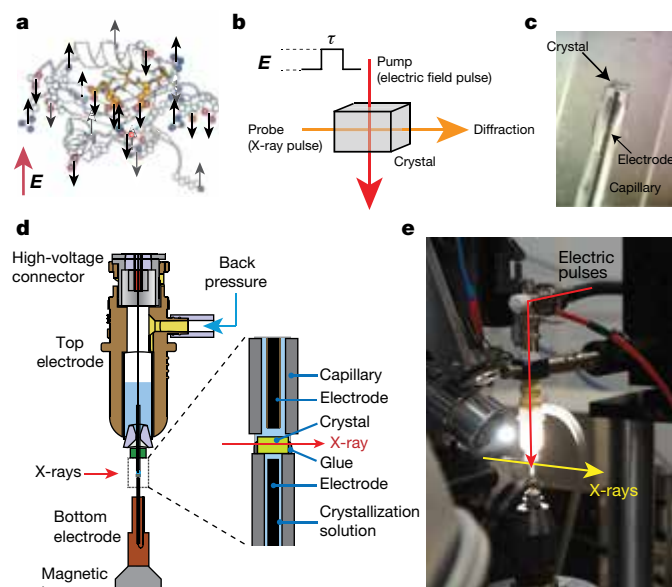


Figure 1 | EF-X principles and implementation. **a**, A sampling of charged residues (in CPK colours) in LNX2^{PDZ2} (Protein Data Bank (PDB) accession 2VWR), exemplifying potential actuators for applied electric fields (E , in red). Ligand is in yellow. **b**, EF-X involves stimulation of motions in protein crystals by an applied electric field (E) of duration τ (the 'pump'), and readout by much faster X-ray pulses (the 'probe'). **c**, An LNX2^{PDZ2} crystal mounted across the orifice of a glass capillary, filled with crystallization solution and a metal electrode. The crystal is sealed onto the capillary by an electrically insulating glue. **d**, The crystal is mounted on the bottom electrode and the high voltage is delivered from a top electrode through a liquid junction composed of crystallization solution. Controlled back pressure on a reservoir of solution in the top electrode keeps the crystal continuously hydrated. **e**, A view of the assembled experimental apparatus.

the ground state, and that forces imposed by $\sim 1 \text{ MV cm}^{-1}$ electric fields represent an effective strategy to bias and expose these states.

Practically, there are several experimental complications (Supplementary Information IA). Of these, the main one is crystal heating caused by electric-field-induced flow of ionic currents through solvent channels. If sufficiently large, this effect leads to dielectric breakdown, arcing, destruction of the crystal, and a dramatic end to the experiment (Supplementary Video 1). However, calculations with estimated conductivities of protein crystals²¹ and typical crystallization solutions suggested that electric fields on the order of 1 MV cm^{-1} should be tolerated for pulse durations up to microseconds (Supplementary Information IA). Together with rise-time limits of our current high-voltage system ($\sim 10 \text{ ns}$) and electrode design, this defines a window of timescales for these experiments at present (Extended Data Fig. 1a). These limits can be extended through further technical development.

On the basis of these considerations, we built a custom setup for room-temperature X-ray diffraction of protein crystals under strong electric field pulses on the sub-microsecond timescale (Fig. 1c–e, Methods and Extended Data Fig. 1). Protein crystals are sandwiched between two glass capillaries filled with crystallization solution and containing metal wires that serve as electrodes (Fig. 1c). The crystal is fixed by an electrically insulating glue to the bottom (ground) electrode and the high-voltage pulse is introduced from a top electrode through a liquid contact with the crystal (Fig. 1d and Supplementary Video 2). See Methods and Extended Data Fig. 1 for design details. This electrode system was integrated into a synchrotron X-ray facility designed for time-resolved crystallography (BioCARS²², Advanced Photon Source; Fig. 1e).

Application of EF-X to the PDZ domain

As an initial model system, we chose the second PDZ domain of the human E3 ubiquitin ligase LNX2 (LNX2^{PDZ2})²³ (Fig. 2a). PDZ domains

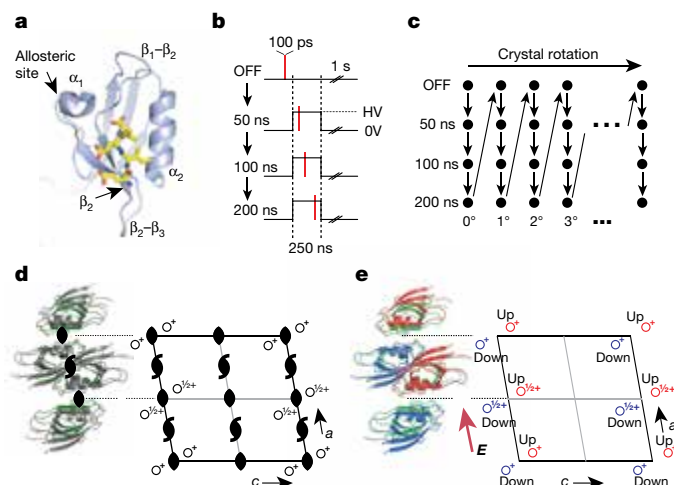


Figure 2 | An EF-X experiment in the LNX2^{PDZ2} domain. **a**, LNX2^{PDZ2} binds target ligands (in yellow) in a groove between the β_2 and α_2 segments. The binding site is coupled to allosteric sites on the β_2 – β_3 segment and the α_1 – β_4 segment (through the β_1 – β_2 loop and α_1 helix). **b**, Data collection involves four sequential X-ray exposures for each crystal orientation: no voltage (OFF), and three time delays (50, 100, 200 ns) after onset of the voltage pulse. One second is allowed between pulses for crystal cooling. HV, high voltage. **c**, The protocol in **b** is repeated for a series of crystal rotations to collect a full diffraction data set. **d**, LNX2^{PDZ2} crystallizes in the C2 space group, which includes two kinds of rotational symmetry (black symbols); this results in four molecules per unit cell and one molecule per asymmetric unit. **e**, With the electric field E (applied along the a dimension), all rotational symmetry is broken. This results in a new unit cell with two molecules per asymmetric unit (red and blue)—one experiencing $+E$, and one experiencing $-E$.

are 90–100-residue proteins that generally bind the C termini of target proteins between the α_2 helix and β_2 strand²⁴. Previous data demonstrate the existence and functional relevance of allosteric coupling of the ligand-binding site to a few distant surfaces²⁵, especially the α_1 helix^{26,27} and the β_2 – β_3 loop²⁸ (see Supplementary Table 1). Otherwise, LNX2^{PDZ2} is a typical protein, with no special features that compromise the generality of this study. Specifically, LNX2^{PDZ2} has no known functional voltage dependence, providing a test that EF-X can be generically used in the context of randomly available formal and partial charges for analysis of protein mechanics.

We performed EF-X experiments on LNX2^{PDZ2} with voltage pulses of 5–8 kV to 50–100- μm -thick crystals, resulting in field strengths of ~ 0.5 – 1 MV cm^{-1} . The pulse durations ranged from 50 to 500 ns, and diffraction was collected with single 100 ps X-ray pulses. The pulse protocol permits us to examine the atomic structure before the electric pulse (voltage-OFF data set) and at any specified time delay after initiation of the electric pulse (voltage-ON data set) (Fig. 2b, c). The OFF data set provides a reference structure for study of electric-field-induced effects. As predicted by our calculations, LNX2^{PDZ2} crystals (and other protein crystals) readily tolerated hundreds of 100–500 ns electric field pulses of $\sim 1 \text{ MV cm}^{-1}$ and X-ray pulses without substantial loss of diffraction (Supplementary Table 2 and Extended Data Fig. 2). We collected a time series from a single LNX2^{PDZ2} crystal, consisting of an OFF data set and ON data sets at 6 kV and at 50, 100 and 200 ns delays from the rising edge of the electric field pulse (Fig. 2b, c and Supplementary Table 3); the variability in timing is less than 1 ns and is therefore negligible given the timescale of this experiment.

Breaking symmetry

An important analytic tool comes from understanding how the electric field affects the symmetry S of the crystal lattice. In general, the unit cell of a protein crystal can be constructed from a set of symmetry operations $\{S\}$ —combinations of translations and rotations—that define its

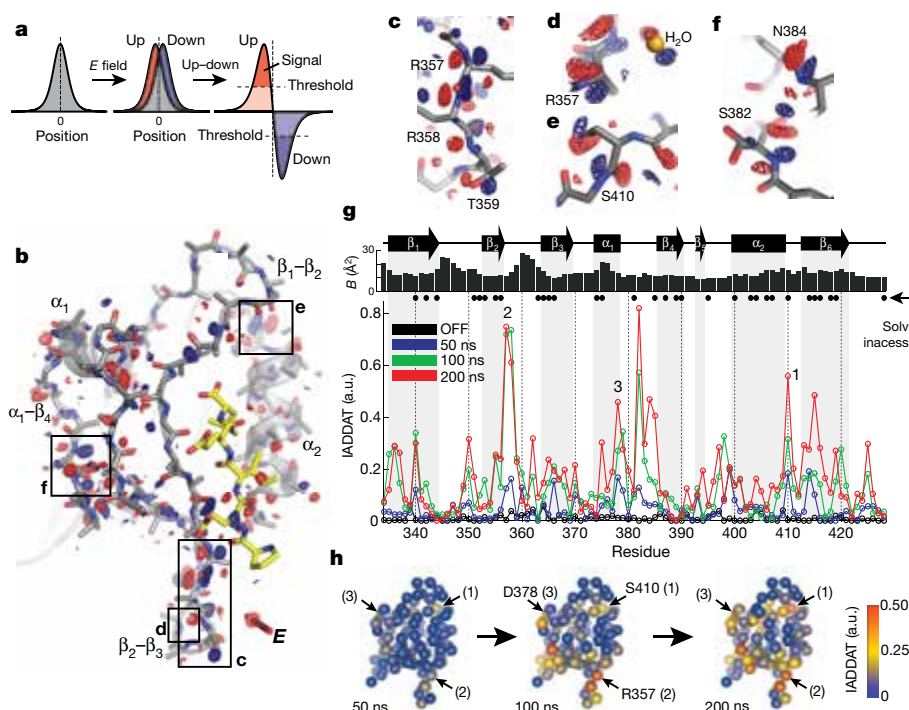


Figure 3 | The up-down internal difference analysis. **a**, In the simplest case, the electric field will shift the electron density distribution for an atom in the up and down molecules (red and blue, respectively) in opposite directions around its centroid in the voltage-OFF molecule (grey) (left and middle). Subtracting the up and down densities and applying a noise threshold (right), we expect peaks of positive (red) and negative (blue) difference density surrounding an atom in the OFF state—the hallmark of an electric-field-induced motion. **b**, The up-down internal difference map for a ‘front’ view of LNX2^{PDZ2}, with regions highlighted in c–f boxed. The red three-dimensional arrow indicates the direction of the electric field, and bound ligand is in yellow. Maps are contoured at +3.5 (red) and –3.5 (blue) σ_{OFF} and, for clarity, are displayed within

a 1.8 Å shell around main chain + C β atoms (see PyMOL session S1 for full map). **c–f**, Examples of electric-field-induced motions—opposing red and blue density—for main-chain, side-chain and solvent atoms. **g**, The IADDAT for the up-down molecules as a function of LNX2^{PDZ2} primary structure and time (blue, green and red traces). The OFF difference density (black) indicates the noise in the analysis. The graphs above indicate buried residues (solvent accessibility <0.15 (Solv. inaccess.)) and refined isotropic B -factor for the voltage-OFF model. a.u., arbitrary units. **h**, The time evolution of the electric-field-induced effects mapped on the tertiary structure of LNX2^{PDZ2}. Spheres indicate C α positions and colours IADDAT.

characteristic space group. For example, the LNX2^{PDZ2} crystals have space group $C2$, which, in addition to translational symmetry, has two kinds of rotational symmetry elements (Fig. 2d). As a consequence, there are four symmetric LNX2^{PDZ2} monomers per unit cell. What happens if an electric field is applied in a certain direction? Clearly, protein molecules in the crystal lattice with different orientations relative to the field will undergo different changes and will no longer be symmetric. The general rule is that any crystal symmetry operator S that does not preserve the orientation of the electric field E will be violated (‘broken’: $S \cdot E \neq E$). For the LNX2^{PDZ2} experiment, the electric field breaks all the $C2$ rotation symmetry operators (Fig. 2e). Now, the four LNX2^{PDZ2} molecules in the unit cell are no longer equivalent, and symmetry is reduced such that two molecules see the field in one direction (we will refer to these molecules as ‘up’), and two see the field in the opposite direction (the ‘down’ molecules). In essence, if the up molecule experiences +6 kV, the down molecule experiences –6 kV, and so the force acting on otherwise equivalent atoms in these structures is opposite in direction. Although it need not be strictly symmetric, we would naively expect this to cause an opposite motion of atoms from their mean positions in the OFF state (Fig. 3a).

This breaking of symmetry provides a powerful way to study the effect of the electric field on the protein structure. We can compare the up and down molecules within the unit cell, an internally controlled experiment that isolates the effect of the electric field on atoms. In contrast, artefacts due to radiation damage and heating are insensitive to the direction of the electric field and cancel out in this analysis (see Methods). In crystallographic terms, we compute an internal difference Fourier map in which we subtract the up and down electron densities

(Fig. 3a). In such a map, the hallmark of an electric-field-induced structural effect is to see peaks of opposite sign around the position of an atom in the voltage-OFF state (red and blue, Fig. 3a).

The up-down map shows pervasive evidence of electric-field-induced atomic motions (Fig. 3b–f). Just as proposed, we observe shifts of backbone, side-chain and solvent atoms in opposite directions between the up and down molecules (Fig. 3c–f). The structural response is distributed broadly over the protein tertiary structure, in both core and surface sites, with some of the strongest signals around the β_2 – β_3 , α_1 – β_4 and α_2 – β_6 segments (Fig. 3b). To examine the response quantitatively, we integrated the absolute difference electron density above a noise threshold (IADDAT) in a volume shell around the protein backbone²⁹ (Fig. 3g). The up-down effect in the OFF state provides a measure of noise (black trace, Fig. 3g). By comparison, we observe a robust signal in the presence of the electric field that evolves over time from 50 ns to 200 ns (Fig. 3g, blue, green and red traces, and Fig. 3h). The electric-field-induced motions do not simply reflect solvent exposure or thermal (B) factors related to positional disorder (for all cases, $P > 0.1$, Fisher Z-test; see Methods) (Fig. 3g). Many of the affected residues do not have formally charged side chains, indicating that they move due to local dipoles or due to structural coupling with other charged residues. An extensive statistical validation of signal to noise is presented in Extended Data Figs 3a–c and 4, Supplementary Tables 4–6 and Supplementary Information IB.

The signal evolves heterogeneously over the structure (Fig. 3g, h), with some regions moving over the full time period (for example, peaks 1 and 3), and others complete at intermediate times (for example, peak 2). This variation in characteristic timescales of motion in different

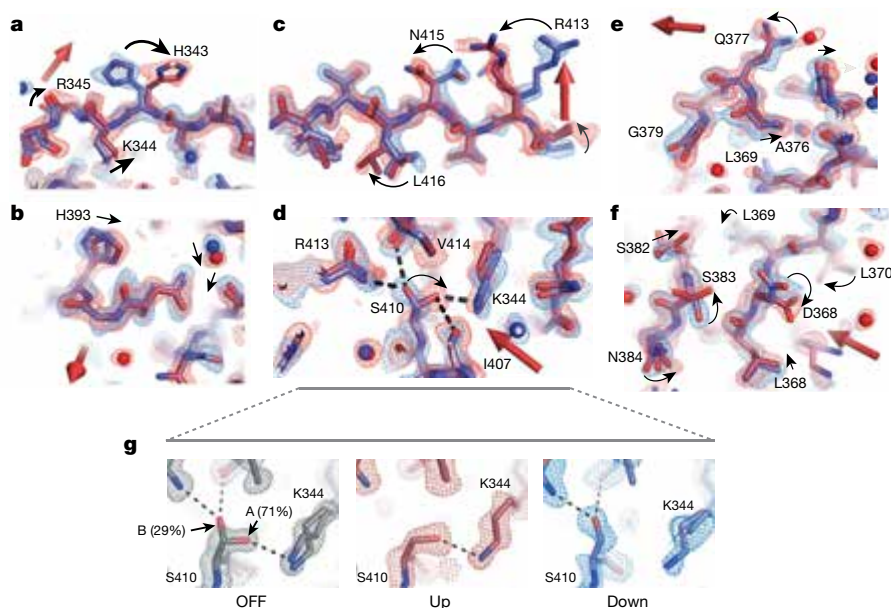


Figure 4 | A gallery of electric field-induced structural effects.

a–f, Refined models and associated $2F_o - F_c$ electron density contoured at 1.5σ for the up (red) and down (blue) LNX2^{PDZ2} structures (6 kV, 200 ns delay); the direction of the electric field is indicated by the three-dimensional arrow. The data show examples of rotamer flips (H343; **a**), continuous displacements (H393; **b**), potentially coupled rotamer flips (R413, N415, L416; **c**), rearrangements of hydrogen bonding (S410; **d**), motions of secondary structure elements (the α_1 helix; **e**), and complex combinations of these effects (**f**). Per sign convention, atoms coupled to a positive charge

in the up model would move in the direction of the field and in the down model, against the field. Motions occur at solvent-exposed (**a–c**, **e**) and -buried (**d**, **f**) regions (PyMOL sessions S2–S4 and Extended Data Fig. 6). **g**, S410 shows partial occupancy in two rotameric states (marked A and B) in a 1.1 Å room-temperature ground-state structure (OFF, Extended Data Table 3). These states are biased by the electric field such that the up and down models each adopt one of the two ground-state configurations (middle and right). Maps are contoured at 1.5σ . See Extended Data Fig. 7 for more examples.

regions of the structure is a property that, with further study, could be deeply informative about the underlying pattern of forces between amino acid residues. A broad analysis of crystal growth conditions, diffraction quality and symmetry suggests that many protein crystals should be amenable to the EF-X experiment, including use of the up–down difference method (Supplementary Information ID).

Modelling of excited states

We refined atomic structures of the up and down states of the LNX2^{PDZ2} domain at 200 ns from the onset of the electric field. Since the field only subtly biases the ground state conformation, we carried out refinement against extrapolated structure factors (ESFs)^{30,31} (Extended Data Table 2). The ground state (OFF model, Extended Data Table 3) was used as a starting point, with progress supported by *R* factors ($\Delta R_{\text{work}} = -6.96\%$, $\Delta R_{\text{free}} = -5.92\%$; Methods, Extended Data Fig. 5 and Supplementary Information IB). Propagation of errors suggests that the ESF structures at 200 ns have an effective resolution of 2.3 Å.

The structures demonstrate electric-field-induced perturbations of nearly every type of physical interaction throughout the protein structure—induction of side-chain rotamer flips (Fig. 4a), continuous displacement of backbone atoms, side chains and bound waters (Fig. 4b), propagated rotamer shifts suggesting collective motions through the structure (Fig. 4c), breaking and re-forming of hydrogen bonds (Fig. 4d), global motions of entire secondary structure elements (Fig. 4e), and complex coordinated changes in large regions (Fig. 4f). Extended Data Fig. 6 shows additional examples. For some residues, the electric field biases the occupancy of pre-existing alternate conformational states in the voltage-OFF structure (Fig. 4g and Extended Data Fig. 7). These residues are differentially forced into either of the alternative configurations depending on the direction of the applied field. Thus, rather than inducing non-physiological states, EF-X appears to expose low-lying conformational states that are energetically near to the ground state.

These data validate the broad goals of EF-X—to globally perturb and record subtle motions in a mechanistically unbiased manner at

atomic resolution. A key feature is the ability to actively populate and directly model the structures of low-lying excited states around the ground states of protein molecules, the configurations most likely to be relevant over the functional reaction coordinate. In addition, the ability to collect data sets at various time delays after the initiation of the electric field pulse means that we can observe these motions as they happen in time and make experimental movies of the temporal evolution of protein motions¹⁷.

The biological relevance of stimulated motions

We asked what the electric-field-induced motions tell us about the biology of the PDZ domain. The backbone motions accumulate in four parts of the protein—the α_1 helix and the β_1 – β_2 , β_2 – β_3 and α_2 – β_6 segments—all known to be functionally coupled to ligand binding (Fig. 5a, b). The partially buried α_1 helix and the α_1 – β_4 surface form the central components of allosteric communication in PDZ domains^{27,32–35}, and residues in these regions undergo systematic electric-field-induced shifts and rotameric transitions (Figs 4f and 5a). In addition, the β_1 – β_2 and β_2 – β_3 segments move and become more ordered (Extended Data Fig. 6d), transitions reminiscent of ligand-induced changes in many PDZ domains^{24,25}. Finally, the electric-field-induced switching of S410 between two hydrogen-bonding networks (α_2 – β_6 region, Fig. 5d, g) positions a conserved buried cationic residue (K344) in the ligand-bound configuration in several PDZ homologues³⁶ (Extended Data Fig. 6e).

To test the relationship of electric-field-induced motions to PDZ function rigorously, we analysed the ligand-induced displacements of main-chain atoms averaged over 11 diverse homologues of the PDZ family (Fig. 5c, e). Ligand-induced motions shared by these homologues are most pronounced in the β_1 – β_2 , β_2 – β_3 and α_2 – β_6 segments, and in the α_1 and α_2 helices (Fig. 5c), comprising most regions with electric-field-induced motions. These regions are also linked by the protein sector^{26,37}—a group of amino acid positions that statistically co-evolves in the entire PDZ family—suggesting that the pattern of ligand-induced motions is an evolutionarily conserved feature in

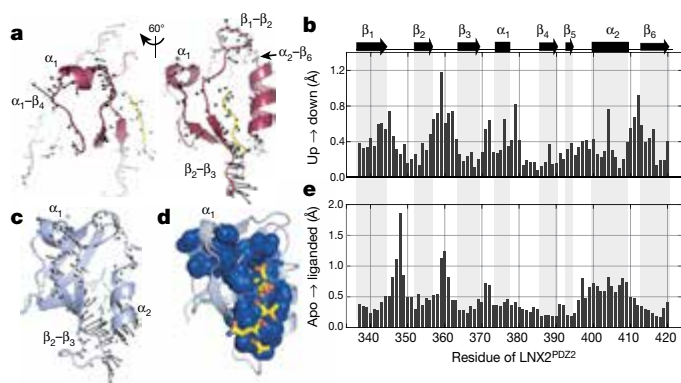


Figure 5 | The relationship between electric-field-induced conformational change and PDZ function. **a**, Two views of the electric-field-induced structural changes in LNX2^{PDZ2} (6 kV, 200 ns data set), with vectors representing the displacements of main-chain atoms transitioning from the up to down models (enlarged $\times 5$ for clarity). The motions are most prominent in the β_1 - β_2 , β_2 - β_3 , α_1 and α_2 - β_6 regions. **b**, The mean displacements of backbone atoms per residue between the up and down states. **c**, Conserved motions of backbone atoms due to ligand binding (apo to liganded) in high-resolution structures of 11 diverse homologues of the PDZ family (vectors enlarged $\times 10$). The motions occur in similar regions as in **a**, but also include the α_2 helix. **d**, The protein sector (blue spheres), a group of coevolving amino acid positions in the PDZ protein family (PFAM 27.0 (ref. 39)); the sector connects the ligand-binding pocket to the β_2 - β_3 segment and to the α_1 - β_4 surface through the β_1 - β_2 loop and the α_1 helix. **e**, The median ligand-induced displacements of backbone atoms per residue (LNX2^{PDZ2} numbering) in the ensemble of 11 PDZ homologues. Statistical comparison with that for the up to down transition (**b**) shows a significant correlation ($P < 0.001$, Fisher Z-test).

the PDZ domain (Fig. 5d). Overall, the pattern of conserved apo to liganded displacements (Fig. 5e) shows a highly significant correlation ($P < 0.001$, Fisher Z-test) with the electric-field-induced up to down motions (Fig. 5b). This result is particularly meaningful because, in principle, ligand binding and electric fields could impose forces in a protein structure in a manner completely distinct from each other, and the comparison reflects an experiment at just one field strength, orientation, and time delay. Thus, EF-X samples motions in the protein structure that are enriched in its biologically relevant mechanical modes.

From structure to mechanics

A central missing tool in our study of proteins is a method to stimulate and record biologically relevant motions over a broad range of time-scales and with atomic resolution. We show that strong but physiological electric fields can be used to examine a wide range of functional conformational changes within a protein. With further development, we expect that EF-X can be broadly used to investigate the structural basis of protein function (see Supplementary Information ID, IE). It will be of interest to extend EF-X to broader timescales of motions (a matter of further engineering, Extended Data Fig. 1a), to characterize motions in proteins with complex multistate conformational changes, and to study the structures of membrane proteins under physiological electric fields.

However, to go beyond the descriptive level of motions to the underlying physics, it is necessary to infer the spatial distribution of forces, and energies, associated with the observed conformational transitions. In this regard, it is informative to compare EF-X with single-molecule force spectroscopy¹⁵. An electric field of 1 MV cm^{-1} (or 10^8 N C^{-1}) exerts 16 pN per elementary charge, a force sufficient to unzip a leucine zipper protein³⁸. Thus, an exciting prospect is to obtain direct force and free energy estimates for both gradual and discrete conformational changes as in force spectroscopy, but with the atomistic detail and temporal resolution made possible by EF-X. This goal is complicated by the cooperative action of amino acids, but EF-X provides a potential

path to address this problem as well. We can collect EF-X data while varying the duration, orientation, spatial pattern and magnitude of applied forces and statistically group residues that move together into collective modes. These modes may represent the basic mechanical units underlying protein function.

This initial report of EF-X does not yet present a simple, turnkey method. Crystal handling, electrode design, data analysis and structure refinement all leave substantial room for improvement. In addition, the analysis of effects induced by one field orientation and at one time-scale is just a starting point for a full description of relevant motions. However, this work provides an experimental foundation for building good physical models for proteins, the critical link between structure and function.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 October 2015; accepted 24 October 2016.

Published online 7 December 2016.

- Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).
- Méndez, J. & Stillman, B. Perpetuating the double helix: molecular machines at eukaryotic DNA replication origins. *BioEssays* **25**, 1158–1167 (2003).
- Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313**, 1638–1642 (2006).
- Noji, H., Yasuda, R., Yoshida, M. & Kinosita, K. Jr. Direct observation of the rotation of F1-ATPase. *Nature* **386**, 299–302 (1997).
- Krishnamurthy, H. & Gouaux, E. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature* **481**, 469–474 (2012).
- Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288**, 88–95 (2000).
- Sprang, S. R. G protein mechanisms: insights from structural analysis. *Annu. Rev. Biochem.* **66**, 639–678 (1997).
- Monod, J., Wyman, J. & Changeux, J. P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).
- Popovych, N., Sun, S., Ebright, R. H. & Kalodimos, C. G. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.* **13**, 831–838 (2006).
- Cooper, A. & Dryden, D. T. Allostery without conformational change. A plausible model. *Eur. Biophys. J.* **11**, 103–109 (1984).
- Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
- Kay, L. E. Protein dynamics from NMR. *Biochem. Cell Biol.* **76**, 145–152 (1998).
- Fraser, J. S. *et al.* Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).
- Sekhar, A. & Kay, L. E. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl Acad. Sci. USA* **110**, 12867–12874 (2013).
- Neuman, K. C. & Nagy, A. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods* **5**, 491–505 (2008).
- Moffat, K. Time-resolved biochemical crystallography: a mechanistic perspective. *Chem. Rev.* **101**, 1569–1581 (2001).
- Ren, Z. *et al.* A molecular movie at 1.8 Å resolution displays the photocycle of photoactive yellow protein, a eubacterial blue-light receptor, from nanoseconds to seconds. *Biochemistry* **40**, 13788–13801 (2001).
- Swain, J. F. & Gierasch, L. M. The changing landscape of protein allostery. *Curr. Opin. Struct. Biol.* **16**, 102–108 (2006).
- Tao, X., Lee, A., Limapichat, W., Dougherty, D. A. & MacKinnon, R. A gating charge transfer center in voltage sensors. *Science* **328**, 67–73 (2010).
- Ben-Chaim, Y. *et al.* Movement of 'gating charge' is coupled to ligand binding in a G-protein-coupled receptor. *Nature* **444**, 106–109 (2006).
- Morozova TYA, *et al.* Ionic conductivity, transference numbers, composition and mobility of ions in cross-linked lysozyme crystals. *Biophys. Chem.* **60**, 1–16 (1996).
- Graber, T. *et al.* BioCARS: a synchrotron resource for time-resolved X-ray science. *J. Synchrotron Radiat.* **18**, 658–670 (2011).
- Rice, D. S., Northcutt, G. M. & Kurschner, C. The Lnx family proteins function as molecular scaffolds for Numb family proteins. *Mol. Cell. Neurosci.* **18**, 525–540 (2001).
- Doyle, D. A. *et al.* Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067–1076 (1996).
- Fuentes, E. J., Der, C. J. & Lee, A. L. Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *J. Mol. Biol.* **335**, 1105–1115 (2004).
- Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).

27. Peterson, F. C., Penkert, R. R., Volkman, B. F. & Prehoda, K. E. Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB–PDZ transition. *Mol. Cell* **13**, 665–676 (2004).
28. McLaughlin, R. N. Jr, Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).
29. Schmidt, M. *et al.* Ligand migration pathway and protein dynamics in myoglobin: a time-resolved crystallographic study on L29W MbCO. *Proc. Natl Acad. Sci. USA* **102**, 11704–11709 (2005).
30. Genick, U. K. *et al.* Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. *Science* **275**, 1471–1475 (1997).
31. Tenboer, J. *et al.* Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* **346**, 1242–1246 (2014).
32. Feng, W., Shi, Y., Li, M. & Zhang, M. Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding. *Nat. Struct. Biol.* **10**, 972–978 (2003).
33. Im, Y. J. *et al.* Crystal structure of GRIP1 PDZ6–peptide complex reveals the structural basis for class II PDZ target recognition and PDZ domain-mediated multimerization. *J. Biol. Chem.* **278**, 8501–8507 (2003).
34. Long, J. *et al.* Supramodular nature of GRIP1 revealed by the structure of its PDZ12 tandem in complex with the carboxyl tail of Frs1. *J. Mol. Biol.* **375**, 1457–1468 (2008).
35. van den Berk, L. C. *et al.* An allosteric intramolecular PDZ–PDZ interaction modulates PTP-BL PDZ2 binding specificity. *Biochemistry* **46**, 13629–13637 (2007).
36. Kang, B. S., Cooper, D. R., Devedjiev, Y., Derewenda, U. & Derewenda, Z. S. Molecular roots of degenerate specificity in syntenin's PDZ2 domain: reassessment of the PDZ recognition paradigm. *Structure* **11**, 845–853 (2003).
37. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
38. Gebhardt, J. C., Bornschlög, T. & Rief, M. Full distance-resolved folding energy landscape of one single protein molecule. *Proc. Natl Acad. Sci. USA* **107**, 2013–2018 (2010).
39. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements R.R. dedicates this paper to Alfred G. Gilman, whose contributions were profound and irreplaceable. We thank the staff at BioCARS, Stanford Synchrotron Radiation Lightsource (SSRL) and the UT Southwestern Medical Center Structural Biology Laboratory for technical support, and D. Borek, C. A. Brautigam, S. Leibler, A. Libchaber, K. Moffat, Z. Otwinowski and members of the Ranganathan laboratory for discussions. R.R. acknowledges support from National Institutes of Health (NIH) grant R01GM123456, the Robert A. Welch Foundation (I-1366), the Lyda Hill Endowment for Systems Biology, and the Green Center for Systems Biology. BioCARS is supported by NIH grant R24GM111072 and through a collaboration with P. Anfinrud (NIH/ National Institute of Diabetes and Digestive and Kidney Diseases). The SSRL is supported by the US Department of Energy (Contract No. DE-AC02-76SF00515) and by the NIH (P41GM103393).

Author Contributions D.R.H. and R.R. conceived the experimental approach. All authors contributed to the experimental design, D.R.H. and K.I.W. built the EF-X apparatus, and D.R.H., K.I.W., M.A.S. and R.R. performed experiments. D.R.H., V.S. and R.R. developed analysis methods and analysed the data. D.R.H. and R.R. wrote the manuscript with input from the other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.R. (rama.ranganathan@utsouthwestern.edu).

METHODS

System design and safety. The design of the experimental system is based on simple physical considerations. An applied electric field E will impose a force on net charge q to cause a displacement Δx along the field. For any residue (or other group of atoms), we can associate a transition dipole moment $\Delta\mu = \sum_i q_i \Delta x_i$ with each motion, where i is an index over atoms (in units of elementary charge times distance (eÅ); $1 \text{ eÅ} \approx 4.8 \text{ D}$). The energetic effect due to the electric field is $-\Delta\mu \cdot \Delta E$, and its significance depends on how it compares to thermal energy $k_B T$; for example, a weakly populated excited state increases in occupancy by ~ 2.7 fold when its energy relative to the ground state is lowered by $1 k_B T$. As shown in Extended Data Table 1, fields of $\sim 1 \text{ MV/cm}$ are in the right range for our purpose. On the basis of this, we designed a $\pm 10 \text{ kV}$ power supply (Spellman HV) which charges a pulse generator (IXYS Colorado), and from which high-voltage (HV) pulses are triggered by a TTL signal from the synchrotron signal processing hardware. This establishes precise timing between X-ray and electric field (EF) pulses. Integrity of the conductive path to the tip of the capillary and its associated propagation delay were determined using an HV probe.

We designed a number of safety features. Custom RG-11 high-voltage cables (Gater Industries) were high-potential tested by the power group at the Advanced Photon Source and were approved for use up to 8 kVDC . EF pulses were generated in 'half-bridge' mode, where residual charge stored by the HV cables after an EF pulse is drained through a large capacitor connected to ground. The counter electrode was designed to avoid any path through air of less than 1 cm to the grounded cable connector exterior. The inhibit feature of the power supply was connected to an interlock system at the beamline facility, ensuring that the system is de-energized upon personnel entry into the beamline hutch. Power supply voltage and the counter electrode backpressure were controlled remotely using a network-connected microcontroller and custom software.

Electrode construction. The RG-11 cable is terminated on one side with an HV connector (LEMO) (for the pulse generator) and on the other side with an SHV connector (for the housing of the top counter electrode) (Extended Data Fig. 1). The housing was prototyped in-house using a three-dimensional printer (Formlabs, Somerville, MA) and custom fabricated commercially (PolyJet technology, PartSnap, Irving, TX). The housing contains a cylindrical glass insert fitted with a silicone gasket and a thin metal wire ($75 \mu\text{m}$ diameter, Cooner Wire, Chatsworth, CA) with a dielectric coating, except at the tip. The wire was guided to the crystal through a glass capillary (0.5 or 1.0λ (140 or $200 \mu\text{m}$, respectively) orifice, Drummond Scientific, Broomall, PA). The electrode housing was filled with crystallization solution and contains a small port (blue arrow, Extended Data Fig. 1b) that allowed for computer-controlled backpressure for slow infusion of liquid through the top electrode to maintain crystal hydration. Bottom electrodes were prepared from glass capillaries (0.25λ , Drummond Scientific) with a $\sim 100 \mu\text{m}$ orifice, cut in half and aminosilanized at the tip surface to improve adhesive capacity. A $75 \mu\text{m}$ diameter uncoated stainless steel wire (Cooner Wire) was threaded until just below this orifice. The capillary was inserted in a reusable goniometer base (MiTeGen, Ithaca, NY) and soaked, in inverted position, in crystallization solution.

Protein expression, purification and crystallization. For LNX2^{PDZ2}, we obtained an expression strain (BL21(DE3)-R3-pRARE2) and plasmid construct (pNIC28-LNX2^{PDZ2}) from the Structural Genomics Consortium (SGC)⁴⁰ (<http://www.thegsc.org>; construct identifier LNX2A-c033). pNIC28-LNX2^{PDZ2} includes residues 336–424 from *Homo sapiens* LNX2, with the F338L mutation described by the SGC, an N-terminal cloning artefact (334–335), and a C-terminal ligand motif Glu-Ile-Glu-Leu (425–428). LNX2^{PDZ2} protein was expressed as an N-terminal hexahistidine fusion in BL21(DE3)-R3-pRARE2 and purified by nickel affinity chromatography (Ni-NTA agarose, Qiagen), cleavage of the TEV tag by 1 U ProTEV per $50 \mu\text{g}$ protein during dialysis into 50 mM HEPES pH 7.5, 500 mM NaCl, 5% glycerol, 0.5 mM TCEP, size exclusion chromatography, and concentrated to 20 mg/ml for storage. Two protocols yielded suitable crystals. In the first, 3.5 mg/ml protein was dialysed twice (12 and 6 h) against 3.15% glycerol, and crystallized by the hanging drop vapour diffusion method in 19% PEG-300, 48 mM citric acid, 35 mM NaH_2PO_4 and 5% glycerol at 20°C . Drops were set up by mixing $0.55 \mu\text{l}$ protein and $1.0 \mu\text{l}$ buffer. In the second protocol, concentrated protein was diluted to 3.5 mg/ml with 10% glycerol and crystallized by hanging drop vapour diffusion in 27 – 31% PEG-300, 43 mM citric acid and 35 mM NaH_2PO_4 at 20°C (drops, $1.0 \mu\text{l}$ protein and $1.0 \mu\text{l}$ well solution).

Crystal mounting. Crystals were manually mounted under a stereomicroscope across the orifice of the pre-soaked bottom electrode, attached to a magnetic goniometer base. Sylgard 184 (Dow-Corning) was prepared to just before full curing and applied around the crystal using a piece of monofilament fishing line (Cajun Line, $0.012''$ diameter, Zebco, Tulsa, OK), taking care to not overcoat the crystal. A MiTeGen polyester sleeve containing $15 \mu\text{l}$ of $50/50$ crystallization solution and water at one end, was slid over the electrode to maintain suitable vapour pressure

for the crystal. The mounted electrode system was placed on the goniometer and the final experimental configuration (Extended Data Fig. 1e, g), was achieved in three steps: (1) coarse relative positioning of the two electrode system using an XYZ translation stage (Thorlabs), (2) cutting the MiTeGen sleeve to expose the crystal, and (3) rapid, camera-guided approach of the top counter electrode until a liquid junction with the crystal was established (Supplementary Video 2).

Data collection and reduction. EF-X data were collected at BioCARS (14-ID) at the Advanced Photon Source, Argonne National Laboratory. The cryostream temperature was set to 289 K , and data were collected using a Rayonix MX340-HS detector with undulators U23 at 10.74 mm and U27 at 15.85 mm (wavelength range of 1.02 – 1.16 Å). The beam size was approximately $90 \mu\text{m}$ (h) \times $60 \mu\text{m}$ (v) and slit settings were $200 \mu\text{m}$ (h) \times $70 \mu\text{m}$ (v). Data collection proceeded in four 180° passes with 4° , 4° , 2° and 1° steps, respectively, and with matching offsets to maximize coverage of reciprocal space (Extended Data Fig. 3 and Supplementary Table 2). Laue data were processed by using Precognition and EpiNorm software, with concurrent processing of OFF and ON frames. The data were integrated to 1.8 Å (Supplementary Table 3) and merged in space group P1 using the C2 unit cell dimensions. The orientation of the imposed electric field relative to the crystal lattice was established directly from indexed diffraction patterns.

Data for the high-resolution room-temperature (277 K) structure (Fig. 4g and Extended Data Table 3) were collected at the Stanford Synchrotron Radiation Lightsource (SSRL, 11-1) using the PILATUS 6M PAD detector from a single crystal and indexed, integrated, scaled and merged in HKL2000 (ref. 41) (HKL Research). The data showed little radiation damage (HKL2000 radiation-damage coefficients of 0.01 – 0.03 ; values >0.1 – 0.15 indicate significant damage⁴²) or non-isomorphism (coefficient 0.001).

Refinement (C2 OFF models). We refined the structure of LNX2^{PDZ2} in the absence of electric field (OFF) first using the high-resolution (1.1 Å) data set collected at SSRL at 277 K , with initial phases obtained by molecular replacement using a cryo structure of LNX2^{PDZ2} (model PDB accession 2VWR). After early simulated annealing, a model was refined by alternating rounds of automated refinement in PHENIX⁴³ and manual adjustments in Coot⁴⁴. Alternate conformations were placed where supported by averaged kick⁴⁵ and $F_o - F_c$ maps. The final model had no Ramachandran outliers. Further refinement yielded a model without alternate conformations, also without Ramachandran outliers (Extended Data Table 3). Initial phases for the 289 K OFF data set collected at BioCARS were determined by direct placement of the high-resolution single-conformer model of LNX2^{PDZ}, with small differences in unit cell dimensions refined by rigid-body refinement in PHENIX. Solvent molecules and alternate conformations were modelled in Coot, with real-space refinement to relieve backbone strain, and limited additional refinement in PHENIX (Extended Data Table 2). Anisotropic displacement parameters were refined only for residues with substantial difference density at atomic positions. Note that for calculation of internal difference maps, it is essential that the model used for phasing be refined in the space group of the OFF crystal lattice to guarantee exact position of symmetry elements. We subsequently expanded the refined model to the asymmetric unit of the reduced-symmetry space group using PDBSET (CCP4 6.4.0)⁴⁶.

Internal difference maps. Difference map Fourier coefficients were calculated directly from merged structure factors using custom MATLAB (Mathworks Inc.) scripts performing the following operations: (1) match structure factors F_{hkl} and $F_{\bar{h}\bar{k}\bar{l}}$ and calculate differences $\Delta F_{hkl} = F_{hkl} - \gamma_{\bar{h}\bar{k}\bar{l}} F_{\bar{h}\bar{k}\bar{l}}$, where $\gamma_{\bar{h}\bar{k}\bar{l}}$ are the correction coefficients for absorption anisotropy derived from OFF data (below); (2) obtain phases of the corresponding structure factors from the C2 OFF model expanded into C1 using PDBSET (CCP4 6.4.0); (3) calculate weights according to

$$w_{hkl} = \left[1 + \frac{\sigma^2(\Delta F)}{\langle \sigma^2(\Delta F) \rangle} + 0.05 \frac{|\Delta F|^2}{\langle |\Delta F|^2 \rangle} \right]^{-1}$$

as previously described^{29,47} and modified⁴⁸ to include a term reducing the contribution of any single structure factor difference. Following Schmidt *et al.*^{29,31}, the difference density maps are improved if structure factors corresponding to large lattice spacings are rejected (here $d_{hkl} > 4 \text{ Å}$), since EF-X typically produces small-scale electron density differences. For anisotropic absorption correction, we compute $\gamma_{\bar{h}\bar{k}\bar{l}} = \tilde{F}_{\bar{h}\bar{k}\bar{l}}^{\text{OFF}} / F_{\bar{h}\bar{k}\bar{l}}^{\text{OFF}}$, where the tilde indicates local scaling⁴⁹ as implemented in SOLVE⁵⁰. Map coefficients were calculated in PHENIX (FFT) with a grid spacing of 0.3 Å . Absolute difference density was integrated in UCSF Chimera²⁷, with calculations based on the C2 OFF model, expanded to C1.

Refinement of excited states. Since the EF breaks C2 symmetry, refinement of the up and down models was carried out in the P1 space group, with the C2 OFF model as a starting point. A P1 unit cell was chosen containing one up and one down chain, requiring a rotation around the c^* axis by $\arctan(b/a)$ (here: 31.1°). To do

this, the C2 OFF model was 'expanded' in PDBSET (CCP4 4.6.0) using symmetry operations (1) X, Y, Z, and (2) (1 - X), Y, (1 - Z) and the resulting model was rotated in PDBSET by the specified angle.

Extrapolated structure factors were calculated as $F^{\text{ESF}} = N(F_0^{\text{ON}} - F_0^{\text{OFF}}) + F_0^{\text{OFF}}$, where $N = 1/(1 - f)$ is the extrapolation factor³⁰. For traditional pump-probe experiments, f is interpreted as the fraction of molecules excited by an optical pulse³⁰; here, it increases the effective population of excited states, facilitating structure refinement. N was chosen as a trade-off between two criteria: map quality, which deteriorates with increasing N , and the appearance of difference electron density peaks consistent with internal difference maps, which initially increases with increasing N (systematic optimization of N in a site-specific manner will be explored in future work). Refinement was performed mostly manually in Coot with determination of R factors in PHENIX, combined with bulk solvent scaling and occupancy refinement every 5–10 modifications (Extended Data Fig. 5a). Near completion, a few rounds of overall coordinate refinement (PHENIX, 10–15 microcycles, small geometric weights) were included. Electron density maps and composite omit maps were calculated in PHENIX with 0.3 Å grid spacing and default settings. Reflections in the R_{free} test set were included in final map calculations. The refinement statistics are given in Extended Data Table 2.

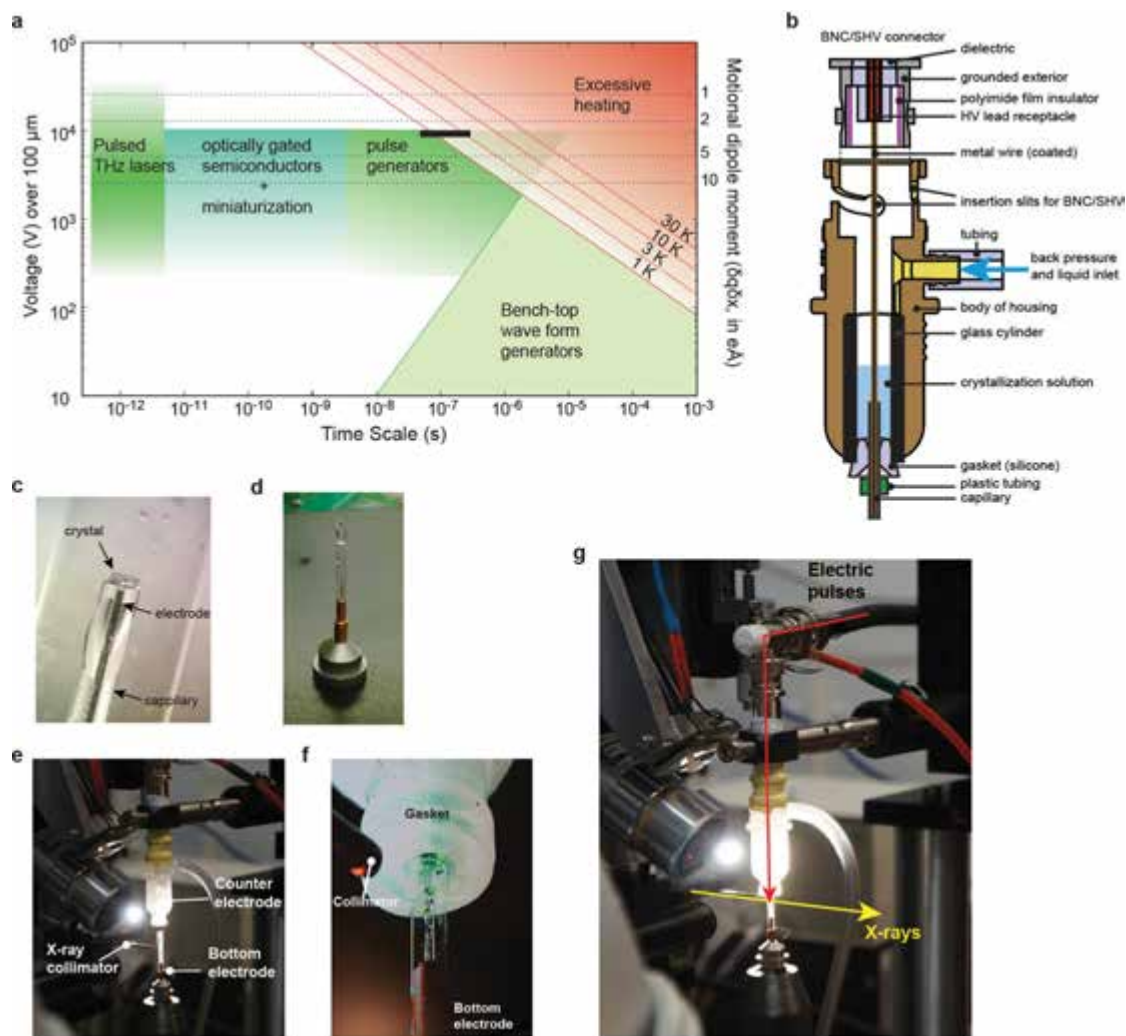
Comparison to homologous PDZ domains. Eleven pairs of high-resolution (≤ 2 Å) X-ray structures of PDZ domains with and without ligand were selected: NHERF-1^{PDZ1}: PDB accessions 1G9O, 1GQ4; PALS-1^{PDZ}: 4UU6, 4UU5; Tiam-1^{PDZ}: 3KZD, 4GVC; ZO-1^{PDZ1}: 4OEO, 4OEP; Erbin^{PDZ}: 2H3L, 1MFG; Dishevelled^{PDZ}: 2F0A, 1L6O; PDZK-1^{PDZ3}: 3R68, 3R69; Shank^{PDZ}: 1Q3O, 1Q3P; GRIP-1^{PDZ6}: 1N7E, 1N7F; PTP-1E^{PDZ2}: 3LNX, 3LNY; PSD-95^{PDZ3} (R.R. *et al.*, unpublished observations). Structures were aligned in PyMOL, using 'super' for backbone atoms, first to the down state of LNX2^{PDZ2}, and then within each pair (Extended Data Fig. 5d legend). For backbone atoms with matching positions in LNX2^{PDZ2}, displacements (Δr) from unbound (apo) to bound (liganded) were then calculated. Atoms with $|\Delta r| < 0.1$ Å were excluded from analysis. Average displacements displayed in Fig. 5c represent the median magnitude and average direction of apo to liganded displacement over homologues.

Statistics. To assess statistical significance of correlations between various experimental measures, the observed quantities were transformed to stabilize variance, reduce kurtosis and approximate a normal distribution. IADAT values (Fig. 3) were square-root transformed, and B -factors (Fig. 3) and displacements (Fig. 5) were log-transformed. To assess the statistical significance of correlations, sample correlation coefficients were then Fisher Z -transformed, and tested for deviation

from a standard normal distribution. For the statistical comparison of the data in Fig. 5b and e, individual residues can be considered independent, yielding $P < 0.001$. More conservatively, one can also take the shorter of the correlation length scales of B -factors and observed displacements (~ 2 residues) as a measure of internal data dependence. This yields a reduced number of independent samples and $P < 0.01$. Thus, the association of Fig. 5b and e is robust to local internal correlations in the data. No statistical methods were used to predetermine sample size.

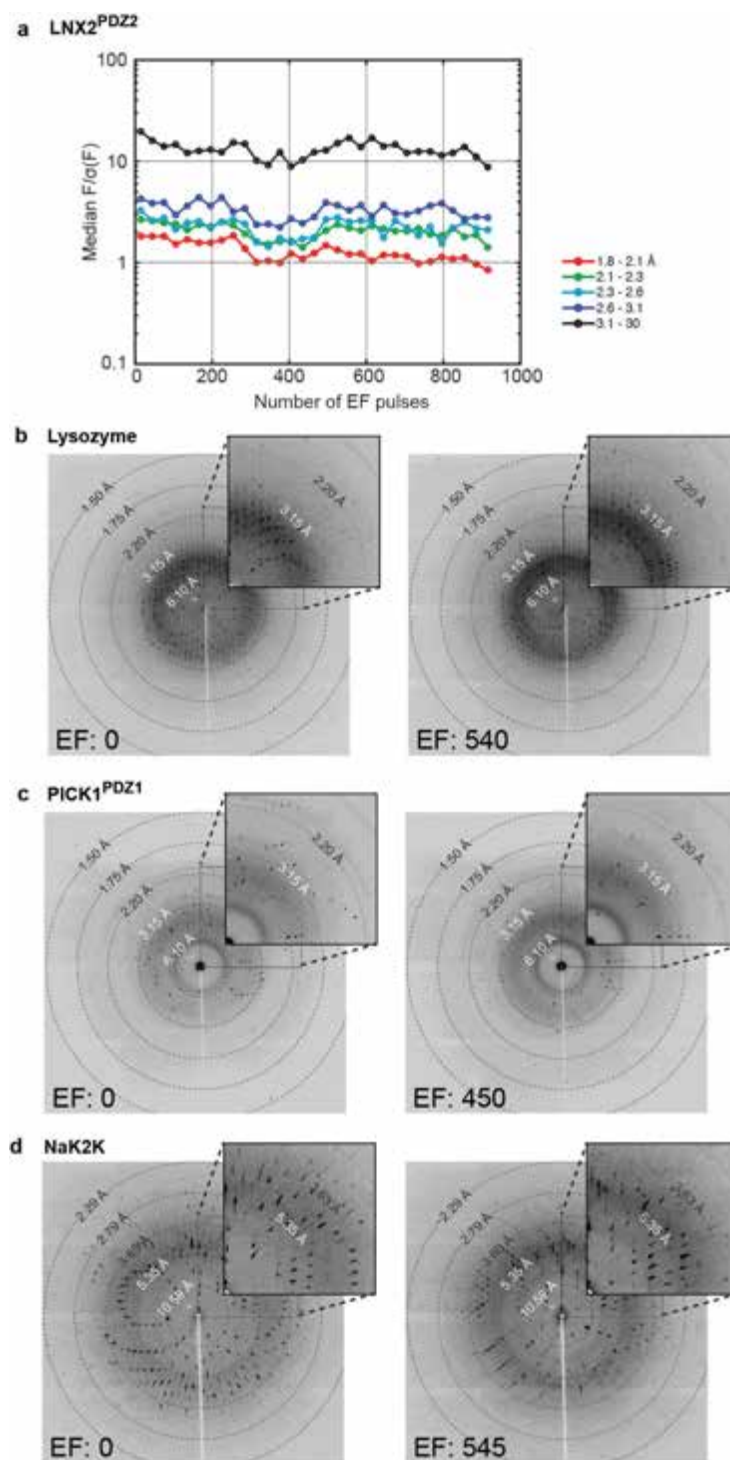
Data availability. Structure factors and refined models have been deposited in the PDB under accessions 5E11, 5E1Y, 5E21 and 5E22.

40. Savitsky, P. *et al.* High-throughput production of human proteins for crystallization: the SGC experience. *J. Struct. Biol.* **172**, 3–13 (2010).
41. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
42. Borek, D., Dauter, Z. & Otwinowski, Z. Identification of patterns in diffraction intensities affected by radiation exposure. *J. Synchrotron Radiat.* **20**, 37–48 (2013).
43. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
44. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
45. Pražnikar, J., Afonine, P. V., Guncar, G., Adams, P. D. & Turk, D. Averaged kick maps: less noise, more signal... and probably less bias. *Acta Crystallogr. D* **65**, 921–931 (2009).
46. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
47. Ursby, T. & Bourgeois, D. Improved estimation of structure-factor difference amplitudes from poorly accurate data. *Acta Crystallogr. A* **53**, 564–575 (1997).
48. Srajer, V. *et al.* Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved Laue X-ray diffraction. *Biochemistry* **40**, 13802–13815 (2001).
49. Matthews, B. W. & Czerwinski, E. W. Local scaling method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. *Acta Crystallogr. A* **31**, 480–487 (1975).
50. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
51. Hoffmann, M. C. Intense ultrashort terahertz pulses: generation and applications. *J. Phys. D* **44**, 083001 (2011).
52. Lefur, P. & Auston, D. H. A kilovolt picosecond optoelectronic switch and Pockel's cell. *Appl. Phys. Lett.* **28**, 21–23 (1976).



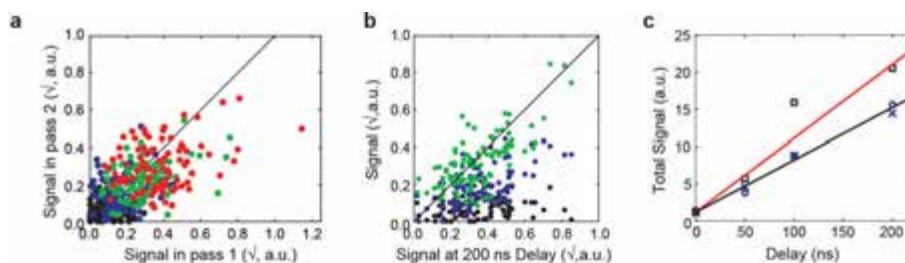
Extended Data Figure 1 | The experimental setup of EF-X. **a**, A plot relating the applied voltage across a 100- μm -thick crystal (left axis) and the size of transition dipole moments of conformational changes that can be excited by $1k_B T$ (right axis) to the duration of the applied electric field. Feasible methods of generating strong electric field pulses are indicated as green and cyan shaded areas. Waveform and pulse generators can provide pulses down to the nanosecond timescale. Faster pulses can be generated using terahertz pulsed lasers with strong electric field components⁵¹ or by optical gating of semiconductors⁵²; such systems are already present at third-generation synchrotron and X-ray free-electron laser facilities. The black bar indicates the approximate range covered by the current experiments. The calculation of temperature jumps caused by the electric field is described in Supplementary Information IA. **b**, Schematic

cross-section of the counter electrode. The blue arrow indicates the path by which backpressure is applied to drive flow through the capillary (see Methods). **c**, Crystals are mounted on top of capillaries containing a metal electrode and soaked in crystallization solution. **d**, The capillary with crystal is mounted in a reusable goniometer base and protected from humidity fluctuations with a polyester sleeve (MiTeGen) containing 50% (v/v) crystallization solution. This assembly forms the bottom electrode. **e**, The counter and bottom electrodes are assembled at the beam line to allow rotation around the capillary axis. **f**, Once the sleeve is trimmed to just above the level of the crystal, the counter electrode is brought in using a translation stage (camera view of the approach) (Supplementary Video 2). **g**, Overview of the final set up with the direction of the X-ray and electric field pulses, reproduced from Fig. 1e.



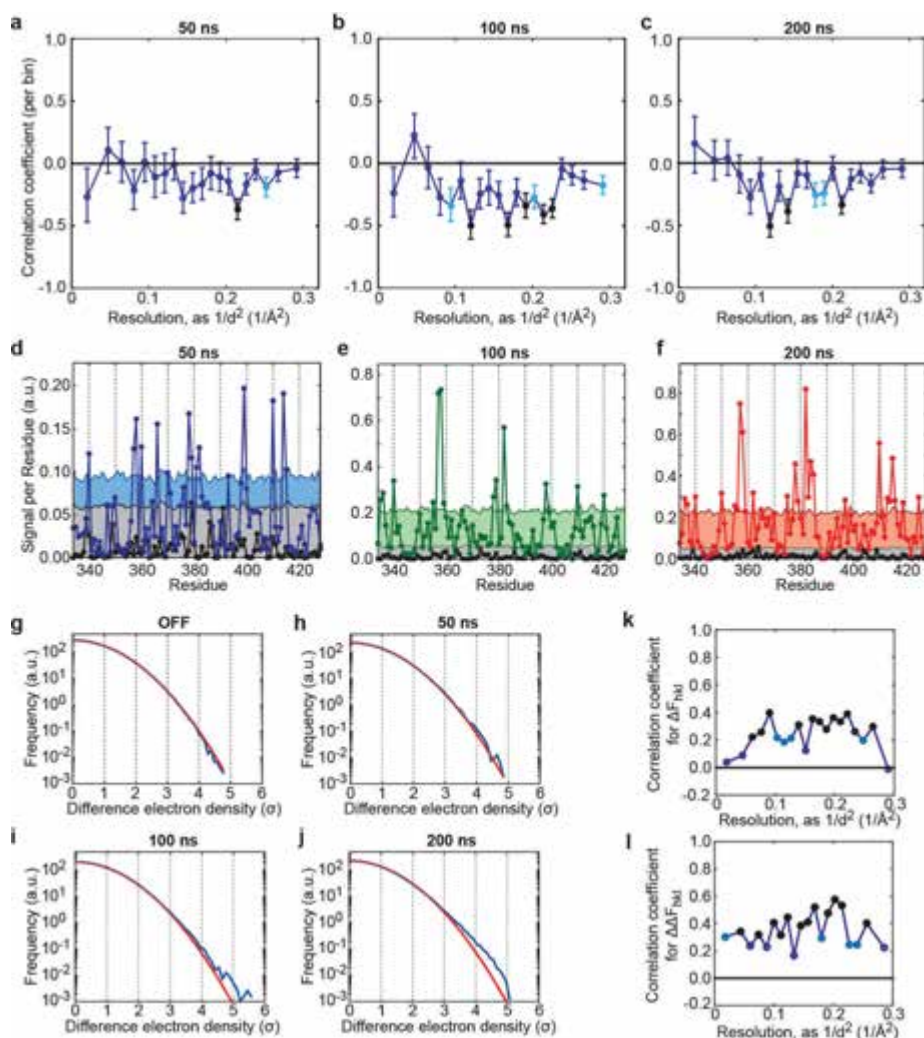
Extended Data Figure 2 | Tolerance of electric field pulses in several protein crystals. **a**, Diffraction quality of a LNX2^{PDZ2} crystal (experiment 3-35, Supplementary Table 2), measured by the ratio of structure factor amplitude to noise ($F/\sigma(F)$) as a function of number of 250 ns, 6 kV electric field pulses and as a function of resolution bin (in colours, see legend). **b–d**, Diffraction images for three other protein crystals before (left) and

after (right) ~500 electric field pulses (precise value indicated). Crystal orientations are different between before and after frames. The data correspond to the following experiments in Supplementary Table 2: **b**, lysozyme, experiment 3-08; **c**, PDZ1 of PICK1, experiment 3-17; **d**, NaK2K, experiment 3-80. The data indicate that several protein crystals can tolerate the EF-X experiment.



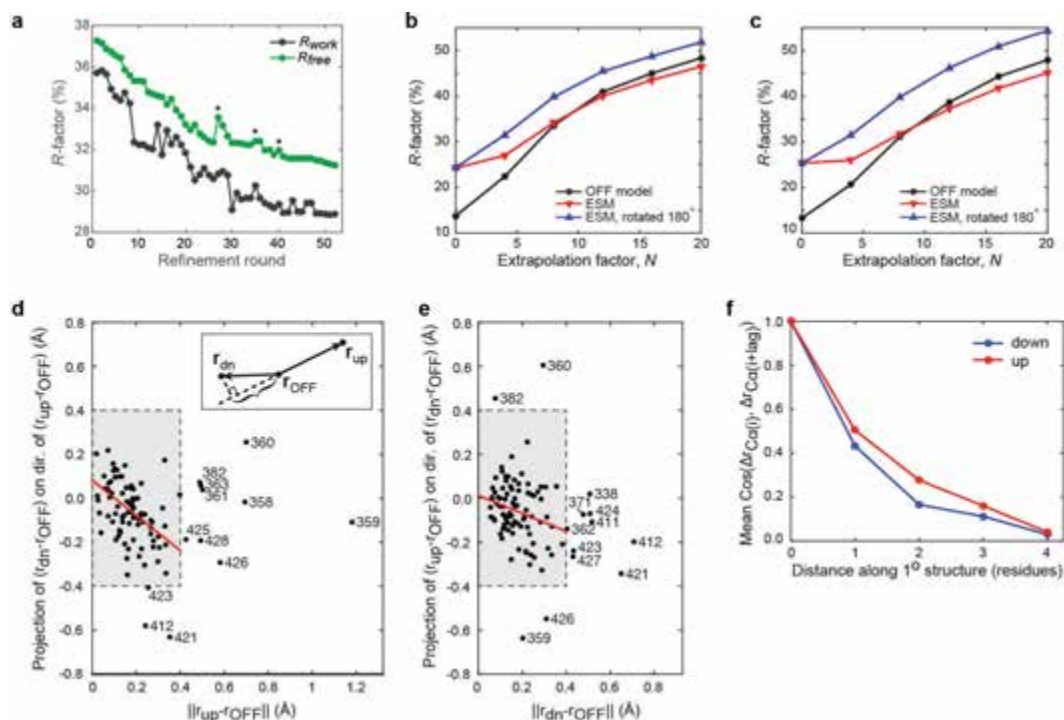
Extended Data Figure 3 | Internal consistency and temporal evolution of internal difference map signal. **a–c**, Analysis for the data presented in Fig. 3. **a**, Consistency of estimated signal per residue derived from two data collection passes on the same crystal (black: OFF; blue: 50 ns; green: 100 ns; red: 200 ns). Overall correlation coefficient 0.59. Signal is defined as the integrated absolute difference density above $2.5\sigma_{\text{OFF}}$ within 1.5 \AA of the protein backbone, square-root transformed to stabilize variance. Per-time-point correlation coefficients are: -0.07 (OFF, $P > 0.1$), 0.23 (50 ns; $P = 0.01$); 0.35 (100 ns; $P < 10^{-3}$) and 0.34 (200 ns; $P < 10^{-3}$).

b, Consistency of the obtained signal per residue between time points. Correlation coefficients are: 0.17 (OFF; $P = 0.05$), 0.55 (50 ns; $P = 1 \times 10^{-9}$) and 0.72 (100 ns; $P < 10^{-20}$). The diagonal is shown for reference. Note that slight correlation in the OFF data set may indicate imperfect correction for anisotropic absorption. **c**, Signal integrated along the entire protein backbone in passes 1 and 2 (blue crosses and circles, respectively) and over the entire data set (squares). The red line indicates a naive expectation of a $\sqrt{2}$ -fold increase in signal-to-noise ratio.



Extended Data Figure 4 | Validation of signal in structure factors and difference maps. **a–c,** A negative correlation between $\Delta F_{hkl} = \Delta F_{hkl}^{\text{ON}} - \Delta F_{hkl}^{\text{OFF}}$ and $\Delta F_{\bar{h}\bar{k}\bar{l}} = \Delta F_{\bar{h}\bar{k}\bar{l}}^{\text{ON}} - \Delta F_{\bar{h}\bar{k}\bar{l}}^{\text{OFF}}$ is consistent with oppositely directed motions in the up and down states. Analysis is performed over 20 resolution bins to allow for statistical testing. Shown are the correlation coefficients per bin between ΔF_{hkl} and $\Delta F_{\bar{h}\bar{k}\bar{l}}$. In a linear response approximation and in the absence of measurement error, we expect $\Delta F_{hkl} = -\Delta F_{\bar{h}\bar{k}\bar{l}}$. Reflections with $|\Delta F_{hkl}| < \sigma(\Delta F_{hkl})$, $|\Delta F_{\bar{h}\bar{k}\bar{l}}| < \sigma(\Delta F_{\bar{h}\bar{k}\bar{l}})$ or $|\Delta F_{hkl}| > 10$ were excluded from analysis, and likewise for $\Delta F_{\bar{h}\bar{k}\bar{l}}$. Results at 50 ns (**a**), 100 ns (**b**) and 200 ns (**c**). To assess significance, each bin was considered statistically homogeneous, with observations considered independent. Bins with significant negative deviation from 0 (after Fisher Z-transform) are indicated as filled circles ($P < 10^{-3}$: black; $P < 10^{-2}$: light blue). Error bars indicate standard errors based on the assumption of a normal distribution after Fisher Z-transform. **d–f,** Statistical significance of Fig. 3g. Comparison of integrated absolute difference density above 2.5σ , within 1.5 \AA of backbone C, N and O atoms ('signal'; see also Fig. 3a). **d,** Comparison of signal in the OFF state and at 50 ns. The grey-shaded area indicates the 0–95th percentile for random sampling from the OFF map at the same probe volume (because the conformation of the backbone changes from residue to residue, the effective probe volume varies along the protein backbone) and threshold. The blue-shaded area indicates the 0–95th percentile for random sampling using the conservative sampling protocol described in test 4 of the statistical validation. **e,** Same analysis at 100 ns. **f,** Same analysis at 200 ns. Note that we were unable to scale all diffraction images at once, and instead scaled the OFF data with each time point separately. We compare each ON data set to the OFF data as scaled with that time point. As a result, there are small differences between the OFF traces in **d–f**.

g–j, Deviations from a normal distribution for internal difference maps. Shown are voxel histograms for internal difference electron density (DED) maps without applied field (OFF) (**g**), and at 50 ns (**h**), 100 ns (**i**) and 200 ns (**j**) of applied electric field. Red lines indicate fits to a normal distribution based on calculated variance. Blue lines are histograms of voxel internal DED values (map grids of 0.3 \AA). Note that by construction, for internal DED maps the positive and negative sides of the histogram are the same, apart from discretization effects. To assess statistical significance of deviations from normality, we sampled C2 asymmetric units (ASUs) at the Nyquist sampling frequency (here, 0.9 \AA). For the OFF map, we find no significant deviations from normality ($P > 0.1$ for the Jarque–Bera test, the Anderson–Darling test, and the Lilliefors test; all using default settings in Matlab). At 200 ns, each test rejects a normal distribution with $P < 0.01$. At 50 and 100 ns, the results of statistical testing depend on how the internal DED map is subsampled: for a single C2 ASU, none of the tests rejects the null hypothesis, but when the same number of points is sampled from two neighbouring ASUs, the Jarque–Bera test rejects normality ($P < 0.01$ at 50 and 100 ns), suggesting limited deviation from normality. **k, l,** Reproducibility of a structural response to electric field. Correlation of data set 2 (see Supplementary Table 7) to the data set presented in the text. On the basis of ordinary differences ΔF_{hkl} (**k**) and internal differences $\Delta \Delta F_{hkl} = (F_{hkl}^{\text{ON}} - F_{hkl}^{\text{OFF}}) - (F_{\bar{h}\bar{k}\bar{l}}^{\text{OFF}} - F_{\bar{h}\bar{k}\bar{l}}^{\text{ON}})$ (**l**), reflections with $|\Delta F_{hkl}| < \sigma(\Delta F_{hkl})$, $|\Delta F_{\bar{h}\bar{k}\bar{l}}| < \sigma(\Delta F_{\bar{h}\bar{k}\bar{l}})$, or $|\Delta F_{hkl}| > 10$ were excluded from analysis, and likewise for $\Delta F_{\bar{h}\bar{k}\bar{l}}$. The standard error of correlation coefficient estimates is ~ 0.07 in **k** and ~ 0.10 in **l**. Each bin is statistically homogeneous and observations are considered independent. Resolution bins with significant positive deviation from 0 (after Fisher Z-transform) are indicated as filled circles ($P < 10^{-3}$: black; $P < 10^{-2}$: light blue).



Extended Data Figure 5 | Refinement, voltage-ON model at 200 ns.

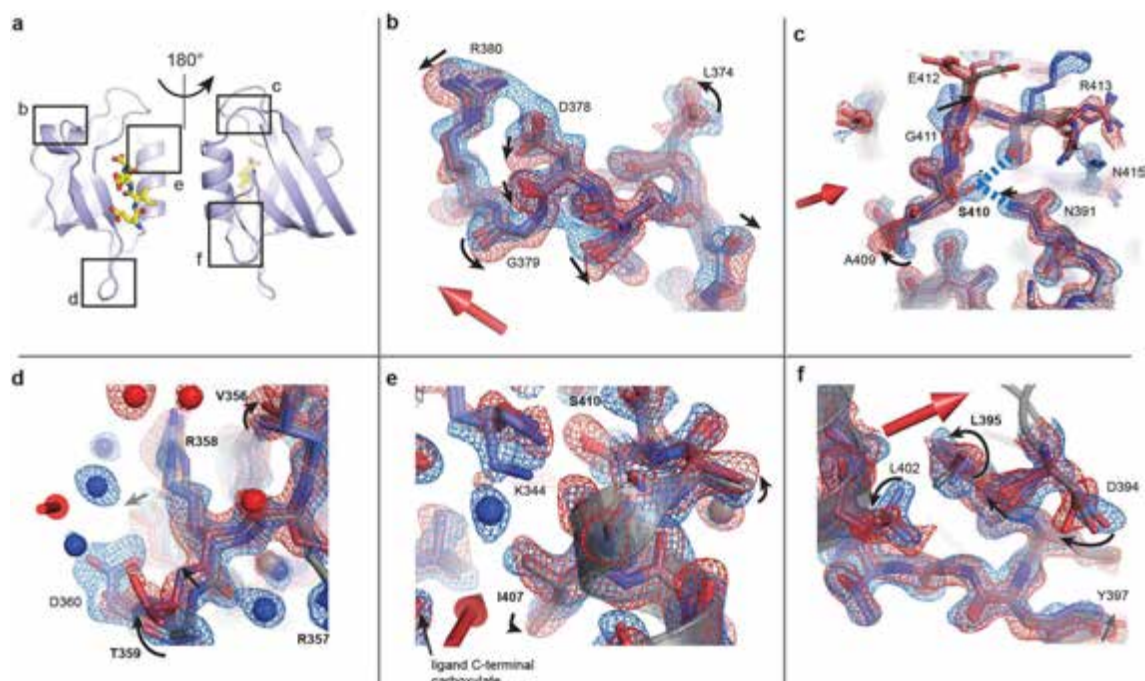
a, Progress of refinement against extrapolated structure factors.

Rounds marked by asterisks involved automated refinement with mild stereochemistry constraints to reduce deviations from optimal geometry due to manual refinement in Coot. Fluctuations in R_{work} appear to be mostly due to the PHENIX bulk solvent scaling calculation used in R factor calculation. **b**, R factor for comparison of extrapolated structure factors, as a function of the degree of extrapolation, N , as derived from data set 2 (150 ns; see Supplementary Table 7), against calculated structure factors (F_c) derived from (1) the OFF model (black), (2) the excited state model (ESM) (red), and (3) an 'upside-down' ESM obtained by 180° rotation around the C2 two-fold rotation axis (blue), all derived from data set 1 (Extended Data Table 2). N relates to the fraction f of OFF signal subtracted as $N = 1/(1 - f)$. No refinement against data set 2 was performed except for bulk solvent scaling. No test set was assigned. **c**, For comparison, the same analysis as in **b**, comparing the OFF model and 200 ns ESM model to the 100 ns data (from the same

crystal). **d–f**, Relationship between $C\alpha$ displacements in the up and down

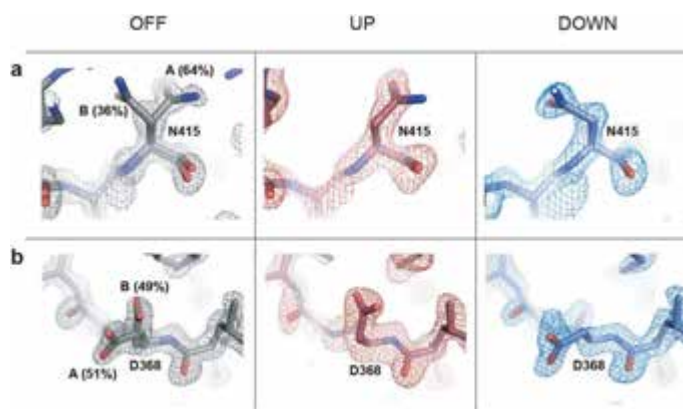
conformations at 200 ns. **d**, **e**, Projection of the down displacement on the direction of the up displacement (all displacements are relative to the OFF model) (**e**); models were superimposed using PyMOL, using C, $C\alpha$ and N atoms of the protein backbone and including only residues 338–356, 362–380, 384–408 and 412–419; this excludes N- and C-terminal regions and mobile parts of the β_2 – β_3 , α_1 – β_4 and α_2 – β_6 loops). Shown are, for example $\Delta r^{\text{down}} \cdot \Delta r^{\text{up}} / |\Delta r^{\text{up}}|$ versus $|\Delta r^{\text{up}}|$, as illustrated in the inset. For small displacements, a simple inverse dependence is expected. This is tested by robust linear regression for (projected) displacements smaller than 0.4 Å (red line fits to data in grey boxes; using default settings in Matlab).

d, Slope = -0.80 ± 0.16 , intercept = 0.081 ± 0.031 Å; correlation coefficient: -0.44 . **e**, Slope = -0.41 ± 0.17 , intercept = 0.012 ± 0.033 Å; correlation coefficient: -0.27 . **f**, Average cosine between displacements of nearby $C\alpha$ atoms as a function of distance along the primary structure.



Extended Data Figure 6 | Additional views of conformational changes due to the electric field. **a**, Reference model indicating regions examined in **b–f**. **b–f**, Maps and models as in Fig. 4, with motions indicated by arrows and residues coupled to ligand binding in PDZ domains shown (as in Supplementary Table 1). **b**, Top view of the α_1 helix, waters omitted and the side chain of Q377 truncated for clarity. **c**, Transverse shift of the α_2 – β_6 loop, and perturbed down state of S410, forming new hydrogen bonds to

R413 and N391 (dashed blue lines). **d**, Upward motion of the β_2 – β_3 loop and change in dynamic disorder of protein and solvent. **e**, Conformational changes at the top of the ligand-binding pocket, with motion of the terminal amine of the K344 towards the ligand carboxylate group in the down state. **f**, Coupled rotameric changes of L402 (α_2 helix), L395 and D394.



Extended Data Figure 7 | Biasing pre-existing conformational heterogeneity in the LNX2^{PDZ2} ground state structure by the external electric field: additional examples. **a, b,** A high-resolution (1.1 Å) room-temperature structure of the voltage-OFF ground state of LNX2^{PDZ2} (Extended Data Table 3), shows partial occupancy of N415 (**a**) and D368 (**b**) in two rotameric states (left). This pre-existing conformational equilibrium is biased in the presence of the electric field (6 kV, 200 ns delay), such that the up and down models each adopt one of the two ground state configurations (middle and right). This supports the result shown in Fig. 4g.

Extended Data Table 1 | Estimates of dipole moments associated with conformational changes

Conformational change	Transition dipole moment (eÅ)	Electric field required for $1 k_B T$ bias (MV/cm)
180° flip of a water molecule ⁴⁶	0.8	3.3
180° flip of a peptide bond ⁴⁶	1.5	1.7
Rotamer flip of a protonated histidine	5.0	0.5
20° turn of a 3-turn α helix dipole ⁴⁶	5.3	0.5
2-ion hop in the KcsA channel ⁴⁹	7.0	0.4
GPCR gating (net, m2R receptor)	~20	~0.13
Gating of a K ⁺ channel ⁵⁰	~100	~0.03

Transition dipole moments were estimated based on the indicated references and for the histidine side chain based on measurements in PyMOL. Shown is the electric field required to bias a conformational equilibrium by $1 k_B T$ with the electric field applied parallel to the transition dipole moment.

Extended Data Table 2 | Data collection and refinement statistics for LNX2^{PDZ2} for EF-X experiment

	LNX2 ^{PDZ2} (OFF)	LNX2 ^{PDZ2} (200 ns)	Extrapolated Differences (8x)
Data collection[†]			
Space group	C2 [†]	P1	P1
Cell dimensions			
<i>a</i> , <i>b</i> , <i>c</i> (Å)	65.30, 39.45, 39.01	38.15, 38.15, 39.01	38.15, 38.15, 39.01
α , β , γ (°)	90, 117.54, 90	113.31, 113.31, 62.28	113.31, 113.31, 62.28
Resolution (Å)	30.08-1.80 (2.0-1.8)*	30.08-1.80 (2.0-1.8)	30.08-1.80 (1.86-1.80)
<i>R</i> _{sym} or <i>R</i> _{merge}	0.088 (0.051)	0.087 (0.053)	n/a
<i>I</i> / σ <i>I</i>	20.7 (37.1)	20.4 (39.9)	6.98 (0.67)
Completeness (%) [‡]	75.1 (42.5)	72.4 (38.1)	70.2 (17.8)
Redundancy	5.8 (3.9)	5.7 (3.6)	n/a
Refinement			
Resolution (Å)	30-1.8 (1.88-1.8)	30-1.8	30-1.8 (1.88-1.8) [§]
No. reflections [¶]	6,565 (288)	11,568	11,291 (288)
<i>R</i> _{work} / <i>R</i> _{free} (%)	13.2/14.8		28.9/31.3
No. atoms (excl. H)	929		1,883
Protein	829		1,712
Ligand/ion	0		6
Water	94		165
B-factors	21.9		16.7
Protein	19.3		16.0
Ligand/ion	n/a		49.5
Water	35.6		22.5
R.m.s deviations			
Bond lengths (Å)	0.020		0.018
Bond angles (°)	1.63		1.80

All data were collected from a single crystal of LNX2^{PDZ2} using Laue crystallography.

*Highest-resolution shell is shown in parentheses. Data reduction in Precognition (Renz Research) differs from conventional data reduction in that weak spots are discarded a priori, resulting in low apparent completeness and high apparent signal and *R*_{merge}, especially at high resolution. Note also that data statistics are reported after global scaling. Subsequent local scaling slightly affects statistics but this scaling mode does not report full last shell statistics. Extrapolated differences were assessed in Xtriage (PHENIX).

†For the purpose of refinement of the OFF model, P1 reflections were merged according to C2 symmetry (merging *R* factor for this: 0.077).

‡See Supplementary Table 3 for data collection and reduction statistics as reported traditionally for Laue crystallography, including assessment of completeness in both C2 and P1. Reported data collection statistics refer to P1.

§Data were retained to the resolution of the two 'parent' data sets (OFF and 200 ns); effective resolution based on propagation of errors is 2.3 Å; completeness over 30–2.3 Å is 89.9%.

||Test sets comprised 5% and 10% of reflections for refinement of the OFF model and refinement against extrapolated structure factors, respectively.

¶The matching number of reflections in the high-resolution shell is coincidental.

Extended Data Table 3 | Data collection and refinement statistics for LNX2^{PDZ2} by room-temperature crystallography

LNX2 ^{PDZ2} (high-resolution)		
Data collection		
Space group	C2	
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	64.91, 39.29, 38.80	
α , β , γ (°)	90, 117.41, 90	
Resolution (Å)	34.45-1.05 (1.05-1.01)*	
<i>R</i> _{sym} or <i>R</i> _{merge}	0.051 (0.54)	
<i>I</i> / σ <i>I</i>	12.84 (0.45)	
Completeness (%) [†]	77.6 (3.0)	
Redundancy	5.8 (1.2)	
Refinement		
	With alternate conformations	No alternate conformations
Resolution (Å)	34.45-1.05	34.45-1.05
No. reflections	35,251 (137)	35,251 (137)
<i>R</i> _{work} / <i>R</i> _{free}	11.9/13.4	12.6/14.0
No. atoms (non-H)	1,039	824
Protein	929	719
Ligand/ion	0	0
Water	104	99
B-factors	19.2	19.7
Protein	17.1	17.3
Ligand/ion	n/a	n/a
Water	37.1	36.2
R.m.s deviations		
Bond lengths (Å)	0.022	0.023
Bond angles (°)	1.86	1.88

On the basis of data collected from a single crystal.

*Highest-resolution shell is shown in parentheses. Data were retained based on CC1/2 > 50%. *I*/ σ *I* falls below 2 at 1.08 Å.

†Completeness over 50–1.5 Å is 98.2%. Completeness falls below 50% (*I*/ σ *I* = 1) at 1.1 Å.

Resolved images of a protostellar outflow driven by an extended disk wind

Per Bjerkerud^{1,2}, Matthijs H. D. van der Wiel^{1,3}, Daniel Harsono⁴, Jon P. Ramsey¹ & Jes K. Jørgensen¹

Young stars are associated with prominent outflows of molecular gas^{1,2}. The ejection of gas is believed to remove angular momentum from the protostellar system, permitting young stars to grow by the accretion of material from the protostellar disk². The underlying mechanism for outflow ejection is not yet understood², but is believed to be closely linked to the protostellar disk³. Various models have been proposed to explain the outflows, differing mainly in the region where acceleration of material takes place: close to the protostar itself ('X-wind'^{4,5}, or stellar wind⁶), in a larger region throughout the protostellar disk (disk wind^{7–9}), or at the interface between the two¹⁰. Outflow launching regions have so far been probed only by indirect extrapolation^{11–13} because of observational limits. Here we report resolved images of carbon monoxide towards the outflow associated with the TMC1A protostellar system. These data show that gas is ejected from a region extending up to a radial distance of 25 astronomical units from the central protostar, and that angular momentum is removed from an extended region of the disk. This demonstrates that the outflowing gas is launched by an extended disk wind from a Keplerian disk.

We obtained high-angular-resolution millimetre-wave observations of the region surrounding the protostar TMC1A using the Atacama Large Millimeter/submillimetre Array (ALMA). We observed the $J=2-1$ rotational transition of the carbon monoxide isotopologues ^{12}CO , ^{13}CO , and C^{18}O . TMC1A is located in the Taurus Molecular Cloud (140 parsecs from Earth), and is a protostellar system with a protostar half the mass of the Sun¹⁴ moving away from the solar neighbourhood (local standard of rest) at a systemic velocity of 6.4 km s^{-1} . It is surrounded by a circumstellar envelope with diameter of around 10^4 astronomical units (AU)¹⁵, has a disk-like structure of radius 200 AU which is inclined 55° with respect to the plane of the sky¹⁴, and has a bipolar outflow extending at least 6×10^3 AU (ref. 16) with position angle $\sim 165^\circ$ east of north. Thus far, TMC1A has been studied at spatial resolutions ranging from several thousand astronomical units down to about 100 AU and the disk is known to exhibit a Keplerian rotation profile¹⁴ at radial distances of about 60–100 AU. The outflow, directed perpendicular to the disk, is bipolar in nature, but is most prominent on the north side of the disk^{16–18}.

The observations (see Methods) were taken at a spatial resolution of 6 AU for TMC1A and cover the inner 200 AU of the outflow as well as the disk surrounding the protostar. The ^{12}CO channel maps in Fig. 1 reveal the walls of the outflow cavity, whereas the ^{13}CO and C^{18}O emission follows the structure of the dust continuum emission (see Extended Data Fig. 1) emanating from the $0.05 M_{\text{sun}}$ disk¹⁴ (where M_{sun} is the solar mass) surrounding TMC1A. A non-disk origin for ^{12}CO is suggested by the noticeable spatial shift of the ^{12}CO emission relative to the ^{13}CO , C^{18}O , and dust continuum emission (see Methods and Extended Data Fig. 1). The morphology of the ^{12}CO emission changes substantially with velocity and small-scale structure (knots) is visible in the maps (Fig. 1). These knots could represent density fluctuations

in the flow^{19,20}, but additional observations at multiple epochs are needed to constrain their nature. The eastern cavity wall (to the left in all figures) of the blueshifted outflow¹⁸ is detected above the 3σ level at velocities offset by more than 2 km s^{-1} from the systemic velocity, and it is clearly offset from the disk and central outflow axes indicated by the dashed lines (Fig. 1). It extends to vertical distances of more than 100 AU from the disk plane, and its direction is consistent with lower-resolution observations of TMC1A tracing 1,000–5,000 AU scales¹⁶. The western cavity wall of the red-shifted outflow is also detected, but at lower velocities compared to the source velocity and at a low signal-to-noise ratio. The northwestern and southeastern cavity walls are not detected, which implies that the radial velocities of these two components coincide with the velocity of foreground material, and therefore suggests that the outflow is rotating (at $v_\phi < 4\text{ km s}^{-1}$; see Methods). Alternatively, to explain their absence, the density and temperature in these regions would have to be much lower than in the two other cavity walls (see Methods), but this is unlikely given the intrinsic bipolar nature of protostellar outflows.

Visual inspection of the channel maps in Fig. 1 suggests that the observed outflowing gas is not launched from within a fraction of an astronomical unit from the central protostar, as would be the case in an X-wind or stellar wind scenario. The corresponding Keplerian radii (plus symbols in Fig. 1) are well outside 1 AU for each channel map. At velocities larger than about 5 km s^{-1} with respect to the systemic velocity, almost no emission is detected along the central outflow axis. It is also clear that outflowing gas is present at large radial separations ($r \approx 50\text{ AU}$) from the central star and close to the disk surface. This is not easily reconcilable with a pure X-wind scenario because the wide-angle flow streamlines predicted by such a mechanism do not correspond with the observed outflowing gas with similar outflow speeds and radial separations, but a range of heights above the disk (see, for example, figure 2 of ref. 21). Consequently, the observed emission cannot be understood using a pure entrainment explanation. The channel maps also show that lower-velocity gas is present at larger distances from the central outflow axis than is higher-velocity gas. This onion-like layered structure⁸ is consistent with observations on larger scales¹⁸. We estimated the outflow launching radii⁸ (footpoint radii, r_0) using two different methods.

First, we fitted a first-order polynomial through all the pixels above the disk midplane in the ^{12}CO channel maps, weighted by the flux density in each pixel (see Methods). This linear least-squares fit provides a direct and straightforward estimate of the footpoint radius for each velocity channel, assuming the gas travels along straight lines (see Methods). The best-fit results, presented in Fig. 2, reveal a range of footpoint radii between about 6 AU and about 22 AU, with a trend where r_0 decreases with increasing velocity. Second, we applied steady-state magnetohydrodynamic wind theory to derive the footpoint radii²². Since the same launching mechanism is responsible for the transfer of both angular momentum and kinetic energy into the wind, both in

¹Centre for Star and Planet Formation, Niels Bohr Institute & Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen K, Denmark. ²Department of Earth and Space Sciences, Chalmers University of Technology, Onsala Space Observatory, 43992 Onsala, Sweden. ³ASTRON, the Netherlands Institute for Radio Astronomy, Postbus 2, 7990 AA Dwingeloo, The Netherlands. ⁴Center for Astronomy, Institute of Theoretical Astrophysics, Heidelberg University, Albert-Ueberle-Straße 2, 69120 Heidelberg, Germany.

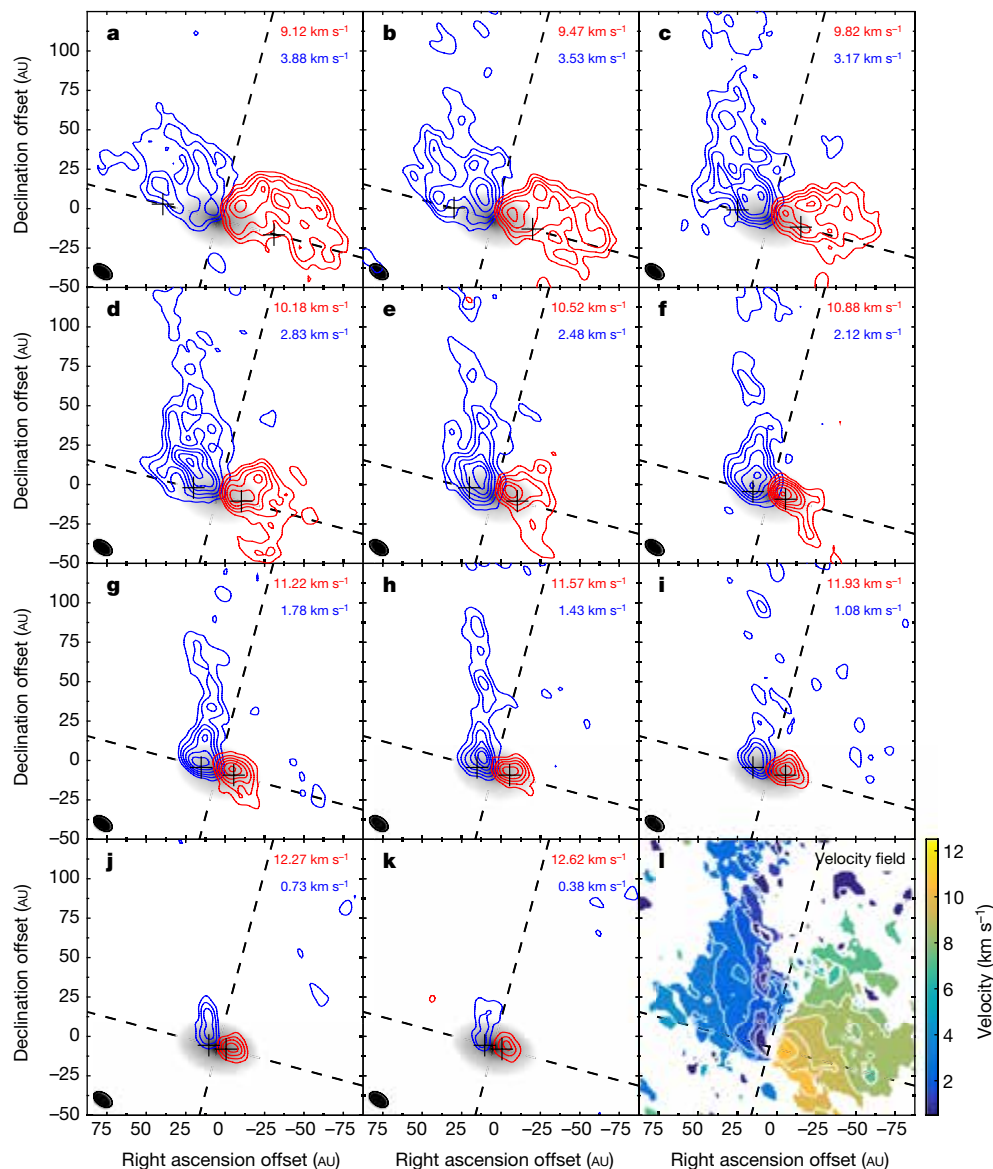


Figure 1 | ^{12}CO channel map of the region. Contours are from 3σ in steps of 1.5σ ($\sigma = 0.8\text{ mJy per beam}$). Blue and red contours represent the emission that is blueshifted and redshifted, respectively, with respect to the systemic velocity (6.4 km s^{-1}). The central channel velocities, the synthesized beam and the radii for the corresponding Keplerian velocities

(plus symbols) are indicated in each panel (a–k). Dashed lines indicate the directions of the disk and the perpendicular outflow axis. The dust continuum emission is shown in greyscale. **l**, The velocity field of the ^{12}CO emission (moment 1 map).

the case of a disk wind and in the case of an X-wind, the outflow and rotational velocity components must be closely linked. Consequently, the footpoint radius can be determined for each position of the map (see Methods). The analysis shows the same trend as the first method, revealing footpoint radii between about 2 AU and about 19 AU (Fig. 3). Thus, the observed emission is consistent with a scenario where a magnetic wind ejects ions from a radially extended region of the disk (which is observed to be in Keplerian rotation around a central mass of $0.4M_{\text{sun}}$; see Methods and Extended Data Fig. 2), that drags molecular gas along. Indeed, the inferred range of footpoint radii is consistent with a disk wind outflow mechanism, whereas, for an X-wind or stellar wind, the footpoints should be located well inside 1 AU.

In the dust continuum data, the flux density distribution reveals an excess in emission relative to the underlying Gaussian profile. The strength of this feature varies slightly with azimuthal angle (most prominent on the southern side of the disk) but is located at a relatively constant radius of around 20 AU (see Extended Data Fig. 3). It is at present unclear whether this feature is related to the launching mechanism, but we note that the radius, interestingly, is very similar

to the estimated maximum footpoint radius of the flow (Fig. 2). We interpret the observed dust emission excess as the result of a density enhancement (and perhaps an elevated dust temperature) at the edge of the outflow launching region.

We measure the specific angular momentum from the velocity field (deprojected from the line-of-sight velocity with respect to the systemic velocity) to be less than 200 AU km s^{-1} in the outflow and it appears to increase with distance from the protostar (Extended Data Fig. 4). This demonstrates that a substantial amount of angular momentum is removed from an extended region throughout the disk. The specific angular momentum of the outflowing gas is comparable to what has previously been reported¹⁸ for the large-scale disk of TMC1A, that is, 250 AU km s^{-1} . Compared to other sources where large-scale emission is observed^{13,23}, the value is relatively low, however. Using the values of the specific angular momentum and the outflow velocity (deprojected from the line-of-sight velocity with respect to the systemic velocity), we can define a locus in the parameter space shown in figure 2 of ref. 13. That figure provides theoretical predictions for the relationship between these quantities, for different launching scenarios. The

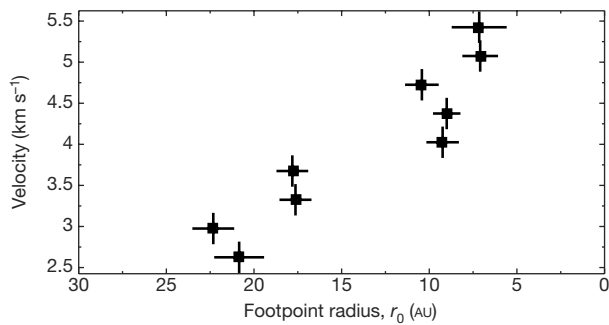


Figure 2 | Direct measure of the disk wind launching point. All pixels above the disk midplane, where the emission flux density is above 3σ , are fitted using a first-order polynomial. The launching point (footpoint radius) is where the best-fit line crosses the disk midplane. Velocity is the absolute value of the observed line-of-sight velocity with respect to the systemic velocity of the protostar (6.4 km s^{-1}). Error bars represent the 2σ confidence interval on the fit and the velocity resolution of the observations. A trend, where higher-velocity gas corresponds to a smaller estimated footpoint radius, is visible.

TMC1A outflow falls in the regime where poloidal outflow velocities are relatively low and launching radii are large. This is consistent with a disk wind launching mechanism but is inconsistent with pure stellar wind or X-wind models.

Observationally, younger outflows are found to be more collimated, have smaller opening angles, and show higher gas velocities than their older counterparts^{24,25}. In this regard, TMC1A does not fall into the category of the very youngest protostars, but rather into the transition period between young and old, where there is still a considerable amount of material available for accretion onto the central protostar. Theoretically, X-winds naturally produce fast, well collimated outflows^{4,5} and stellar winds are effective at spinning down the central protostar⁶, whereas slow outflows and wide opening angles are most easily explained by disk wind models^{7,8}.

The observations presented here demonstrate that the observed TMC1A CO outflow is launched from radial distances substantially displaced from the central protostar, but, since the observations do not resolve the emission on scales below 6 AU, we cannot exclude the possibility of an additional, confined and high-velocity component

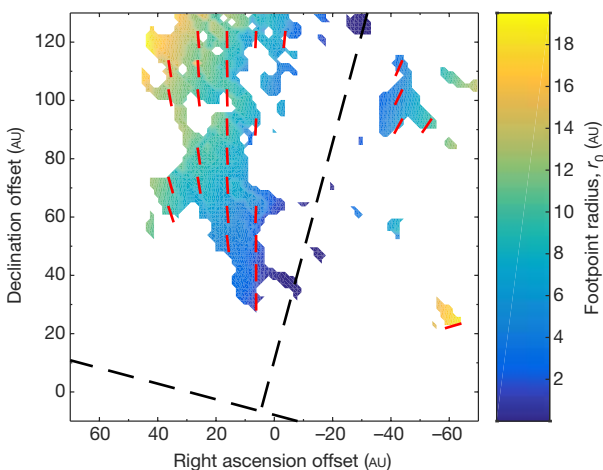


Figure 3 | Disk wind launching point, applying steady-state magnetohydrodynamic wind theory. The footpoint radius is derived under the assumption that the same launching mechanism is responsible both for the angular momentum and kinetic energy extraction into the flow²². All pixels in which the outflow velocity is smaller than the local escape velocity have been masked out. The central outflow axis and the disk midplane are indicated with dashed black lines. For a fraction of pixels, the red dashed lines point to the derived launching point or footpoint; the length of each dash is arbitrary.

not probed by these observations. It has in fact been suggested^{2,26} that a combination of different mechanisms is needed to match all of the observations, within which the disk wind might be important for driving a wide-angle outflow capable of removing a large portion of the infalling envelope²⁴. In general, the most promising theories proposed for protostellar outflow ejection (X-winds, stellar winds and disk winds) have difficulties explaining both large opening angles and bow-shaped structures simultaneously.

A well known observational fact in meteoritics is that part of the chondritic material found throughout the Solar System has a composition consistent with having undergone thermal processing at very high temperatures expected only in the inner Solar System^{27,28}. If the disk wind observed in this work extends to smaller radii (at which the disk wind cannot currently be resolved), it could form the first link in a chain that could transport thermally processed solid material outwards in a protostellar system by allowing it to rain down on the outer part of the disk, whereas an X-wind-type outflow could not²⁹. The TMC1A system has an age of at most a few hundred thousand years³⁰. Although these observations do not probe the very smallest scales, they show that it is possible to drive such a mechanism at times sufficiently early to correspond to the formation epoch of various, chondritic components²⁸ in a young analogue of the Solar System.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 June; accepted 25 October 2016.

1. Snell, R. L., Loren, R. B. & Plambeck, R. L. Observations of CO in L1551—evidence for stellar wind driven shocks. *Astrophys. J.* **239**, L17–L22 (1980).
2. Frank, A. *et al.* Jets and outflows from star to cloud: observations confront theory. *Protostars Planets VI*, 451–474 (2014).
3. Cabrit, S., Edwards, S., Strom, S. E. & Strom, K. M. Forbidden-line emission and infrared excesses in T Tauri stars—evidence for accretion-driven mass loss? *Astrophys. J.* **354**, 687–700 (1990).
4. Shu, F. *et al.* Magnetocentrally driven flows from young stars and disks. I. A generalized model. *Astrophys. J.* **429**, 781–796 (1994).
5. Shang, H., Li, Z. Y. & Hirano, N. Jets and bipolar outflows from young stars: theory and observational tests. *Protostars Planets V*, 261–276 (2007).
6. Bouvier, J. *et al.* Angular momentum evolution of young low-mass stars and brown dwarfs: observations and theory. *Protostars Planets VI*, 433–450 (2014).
7. Blandford, R. D. & Payne, D. G. Hydromagnetic flows from accretion disks and the production of radio jets. *Mon. Not. R. Astron. Soc.* **199**, 883–903 (1982).
8. Pudritz, R. E., Ouyed, R., Fendt, C. & Brandenburg, A. Disk winds, jets, and outflows: theoretical and computational foundations. *Protostars Planets V*, 277–294 (2007).
9. Bai, X. N. Towards a global evolutionary model of protoplanetary disks. *Astrophys. J.* **821**, 80 (2016).
10. Zanni, C. & Ferreira, J. MHD simulations of accretion onto a dipolar magnetosphere. II. Magnetospheric ejections and stellar spin-down. *Astron. Astrophys.* **550**, 99–118 (2013).
11. Bacciotti, F., Ray, T. P., Mundt, R., Eisloffel, J. & Solf, J. Hubble Space Telescope/STIS spectroscopy of the optical outflow from DG Tauri: indications for rotation in the initial jet channel. *Astrophys. J.* **576**, 222–231 (2002).
12. Coffey, D., Bacciotti, F., Ray, T. P., Eisloffel, J. & Woitas, J. Further indications of jet rotation in new ultraviolet and optical Hubble Space Telescope STIS spectra. *Astrophys. J.* **663**, 350–364 (2007).
13. Ferreira, J., Dougados, C. & Cabrit, S. Which jet launching mechanism(s) in T Tauri stars? *Astron. Astrophys.* **453**, 785–796 (2006).
14. Harsono, D. *et al.* Rotationally-supported disks around Class I sources in Taurus: disk formation constraints. *Astron. Astrophys.* **562**, A77 (2014).
15. Chandler, C. J. & Richer, J. S. The structure of protostellar envelopes derived from submillimeter continuum images. *Astrophys. J.* **530**, 851–866 (2000).
16. Yıldız, U. A. *et al.* APEX-CHAMP⁺ high-J CO observations of low-mass young stellar objects. IV. Mechanical and radiative feedback. *Astron. Astrophys.* **576**, A109 (2015).
17. Chandler, C. J., Terebey, S., Barsony, M., Moore, T. J. T. & Gautier, T. N. Compact outflows associated with TMC-1 and TMC-1A. *Astrophys. J.* **471**, 308 (1996).
18. Aso, Y. *et al.* ALMA observations of the transition from infall motion to Keplerian rotation around the late-phase protostar TMC-1A. *Astrophys. J.* **812**, 27 (2015).
19. Ouyed, R., Pudritz, R. E. & Stone, J. M. Episodic jets from black holes and protostars. *Nature* **385**, 409–414 (1997).

20. Hansen, E. C., Frank, A. & Hartigan, P. Magnetohydrodynamic effects on pulsed young stellar object jets. I. 2.5D simulations. *Astrophys. J.* **800**, 41 (2015).
21. Shu, F. H. *et al.* X-winds theory and observations. *Protostars Planets IV*, 789–813 (2000).
22. Anderson, J. M. *et al.* Locating the launching region of T Tauri winds: the case of DG Tauri. *Astrophys. J.* **590**, 107–110 (2003).
23. Lee, C.-F. A change of rotation profile in the envelope in the HH 111 protostellar system: a transition to a disk? *Astrophys. J.* **725**, 712–720 (2010).
24. Arce, H. G. & Sargent, A. I. The evolution of outflow-envelope interactions in low-mass protostars. *Astrophys. J.* **646**, 1070–1085 (2006).
25. Jørgensen, J. K. *et al.* PROSAC: a submillimeter array survey of low-mass protostars. I. Overview of program: envelopes, disks, outflows, and hot cores. *Astrophys. J.* **659**, 479–498 (2007).
26. Banerjee, R. & Pudritz, R. E. Outflows and jets from collapsing magnetized cloud cores. *Astrophys. J.* **641**, 949–960 (2006).
27. Alexander, C. M. O., Grossman, J. N., Ebel, D. S. & Ciesla, F. J. The formation conditions of chondrules and chondrites. *Science* **320**, 1617–1619 (2008).
28. Connelly, J. N. *et al.* The absolute chronology and thermal processing of solids in the solar protoplanetary disk. *Science* **338**, 651–655 (2012).
29. Salmeron, R. & Ireland, T. R. Formation of chondrules in magnetic winds blowing through the proto-asteroid belt. *Earth Planet. Sci. Lett.* **327/328**, 61–67 (2012).
30. Evans, I. N. J. *et al.* The Spitzer c2d legacy results: star-formation rates and efficiencies; evolution and lifetimes. *Astrophys. J. Suppl. Ser.* **181**, 321–350 (2009).

Acknowledgements We thank M. Bizzarro and L. Kristensen for suggestions that improved the paper. This research was supported by the Swedish Research Council through contract 637-2013-472 (to P.B.). M.H.D.v.d.W.

and J.K.J. acknowledge support by a Lundbeck Foundation Junior Group Leader Fellowship as well as the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 646908) through ERC Consolidator Grant 'S4F'. The Centre for Star and Planet Formation is funded by the Danish National Research Foundation. D.H. is funded by the Deutsche Forschungsgemeinschaft Schwerpunktprogramm (DFG SPP 1385) 'The First 10 Million Years of the Solar System—A Planetary Materials Approach'. We also thank the staff at the Nordic ALMA Regional Centre node for assistance with the preparation and calibration of the data. D.H. thanks Leiden Observatory for providing the computing facilities. This paper makes use of ALMA data (see Methods section 'Data availability'). ALMA is a partnership of the ESO (representing its member states), the NSF (USA) and NINS (Japan), together with the NRC (Canada), the NSC and ASIAA (Taiwan), and the KASI (South Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by the ESO, AUI/NRAO and NAOJ.

Author Contributions P.B. and M.H.D.v.d.W. led the project and were responsible for the data reduction, analysis and writing of the observing proposal and manuscript. D.H., J.P.R. and J.K.J. contributed at various stages to the data reduction and analysis, discussed the results and contributed to the proposal and manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.B. (per.bjerkeli@bjerkeli.se).

Reviewer Information *Nature* thanks Y. Aso, D. Coffey and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

ALMA observations and data processing. TMC1A was observed with ALMA on 2015 October 23 and 30. The observations presented here are part of the Cycle 3 programme 2015.1.01415.S. The primary beam of the ALMA 12-m dishes covers a field of 22 arcsec in diameter around TMC1A and the source was observed in the $J=2-1$ rotational transitions of ^{12}CO , ^{13}CO and C^{18}O at 230.5 GHz, 220.4 GHz and 219.6 GHz, respectively. 49 antennas in the 12-m array were used during the observations providing baselines in the range 85 m to 16,196 m. The CO observations were carried out at a spectral resolution of 244 kHz (0.32 km s^{-1}) and the total bandwidth is 117 MHz. In addition to these basebands, one baseband was used for continuum observations where the spectral resolution was set to 976 kHz (1.35 km s^{-1}) and the total bandwidth was 1,875 MHz. The precipitable water vapour during the observations was between 0.28 mm and 0.63 mm. The phase centre of the observations was right ascension $\alpha_{2000} = 04 \text{ h } 39 \text{ min } 35.2 \text{ s}$, declination $\delta_{2000} = +25^\circ 41' 44.27''$. The peak of the continuum emission is slightly offset from this coordinate, that is, by $+0.03''$, $-0.05''$ ($+5 \text{ AU}$, -7 AU corresponding to $\alpha_{2000} = 04 \text{ h } 39 \text{ min } 35.2 \text{ s}$, $\delta_{2000} = +25^\circ 41' 44.23''$). The X-ray source J0440+2728 was used as phase calibrator and the blazar J0510+1800 was used as bandpass calibrator. The flux calibration uncertainty is estimated to be less than 10%.

The data calibration and imaging was carried out in CASA³¹ (version 4.5.0) and followed standard procedure. The continuum is subtracted in the Fourier plane (uv domain) by fitting a constant to the line-free channels. The calibrated visibilities for the continuum are transformed into the image domain using the CLEAN algorithm³² with Briggs weighting and the robust parameter set to 0.5. For the line emission, natural weighting is used to provide the highest signal-to-noise ratio. To improve the signal-to-noise ratio, we used a visibility taper at 0.04 arcsec for the continuum and ^{12}CO maps (shown in Fig. 1) and a visibility taper at 0.10 arcsec for the ^{13}CO and C^{18}O maps (Extended Data Fig. 1). All spectral line output images have a spectral resolution of 0.35 km s^{-1} . The interferometric nature of the observations leads to spatial filtering of large-scale structures, which, for these observations, leads to a maximum recoverable scale of 0.4 arcsec (about 60 AU at a distance of 140 parsecs). This implies that we do not detect any emission that is extended over scales larger than 60 AU and we do not detect emission at velocities below 2 km s^{-1} relative to the systemic velocity of 6.4 km s^{-1} . Hence, we probe material moving with a velocity offset from any extended emission in the system¹⁸ and we do not recover any foreground emission from the envelope.

Analysis of the continuum and spectral line maps. Each velocity channel is analysed individually using MATLAB. For the presented maps, the first contour is always at 3σ . The root-mean-square level of each map is calculated in a 1.3 arcsec by 1.3 arcsec emission-free region located at a distance of 1.5 arcsec from the continuum peak. The presented data has not been corrected for the primary beam response. This has no effect on the maps presented, since the correction is less than 1% within 2 arcsec of the phase centre of the observations.

Origin of the emission. A crucial part of the analysis is to identify the molecular emission that arises from the disk. This can be done through direct comparison of the ^{12}CO , ^{13}CO and C^{18}O emission. The line ratios between the isotopologues are close to one, suggesting that the medium is optically thick in ^{12}CO . To estimate the optical depths, the emission of the isotopologues at $\nu \approx 9 \text{ km s}^{-1}$ is used. Assuming a kinetic temperature of 100 K and adopting isotopic ratios of 60 and 550 for ^{13}CO and C^{18}O , we calculate the optical depths (τ) to be 0.04, 0.2 and 25 for C^{18}O , ^{13}CO and ^{12}CO , respectively. Even in the line wings, the optical depth of ^{12}CO is much greater than 10. The spatial distribution of ^{12}CO differs noticeably from that of ^{13}CO and C^{18}O ; the latter two roughly trace the rotation of the disk (Extended Data Fig. 1). Furthermore, ^{12}CO is not detected in the outer parts of the disk, whereas ^{13}CO and C^{18}O are. This implies that the ^{12}CO in the disk is invisible. In Extended Data Fig. 1, the extent of the disk is derived from integrating the line wings, avoiding the line centre at $\nu < 2 \text{ km s}^{-1}$. If the ^{12}CO emission indeed arises from the disk, the only way to explain the difference in isotopologue distribution is to have foreground material that blocks out the ^{12}CO emission that lies at $> 2 \text{ km s}^{-1}$ from the systemic velocity, which is unlikely because observations³³ and modelling³⁴ of the ambient cloud indicate a median full-width at half-maximum (FWHM) of 1.2 km s^{-1} . It is thus not entirely clear why no ^{12}CO emission is detected in the disk. However, at large radii, the Keplerian speed approaches 2 km s^{-1} , and thus the emission could be absorbed by the foreground ambient material. A simple power-law disk model is used to estimate the CO isotopologue emission arising from a Keplerian disk. The continuum radiative transfer tool RADMC-3D³⁵ and non-LTE molecular excitation analysis³⁶ are used to simulate the predicted molecular emission. The models were processed through CASA using the sm ('simulation') tool in order to simulate the visibilities given the antenna positions. The models indicate that the observed spatial distribution of ^{12}CO , ^{13}CO and C^{18}O should be copatial in the case of a pure Keplerian disk.

To examine the rotation of the disk on the scales where the outflow is launched, we also fit two-dimensional Gaussian distributions to all individual channel maps for ^{13}CO and C^{18}O to find the peak positions. We exclude the ^{12}CO emission from this analysis owing to the substantial contribution from the outflow. The analysis allows us to conclude that the ^{13}CO and C^{18}O emission close to the launching region is Keplerian in nature down to radial distances of about 20 AU from the protostar. The central mass is estimated at $(0.4 \pm 0.1) M_{\text{sun}}$ (Extended Data Fig. 2), taking the inclination angle into account ($i = 55^\circ \pm 10^\circ$), which is slightly lower than previous estimates^{14,18}, and closer to what is obtained when modelling the emission with a rotating infalling envelope³⁷.

The offset between the northeastern cavity wall and the central flow axis, combined with the non-detection of emission from the northwestern and southeastern cavity walls, suggests that the outflow is rotating. The radial velocity of the absorbed outflow components would fall close to the systemic velocity in the case where the ratio between the outflow and rotation velocity components (ν_{out} and ν_{φ} , respectively) is close to $\tan i$. In the case where ^{12}CO predominantly traces outflowing gas, this naturally also explains why we see redshifted emission close to the disk surface, since this is the region where the trajectory of the gas has not yet reached its asymptotic direction. In the analysis presented in this Letter, we calculate the rotation and the outflow velocity components from the observed line-of-sight velocities, corrected for the systemic velocity and deprojected by the inclination angle of the outflow (assumed to be perpendicular to the disk), that is, $\nu_{\varphi} = \nu_{\text{los}}/\sin i/2$ and $\nu_{\text{out}} = \nu_{\text{los}}/\cos i/2$. We thus assume that the outflow is symmetric and that the rotational velocity is constant along the outflow at the scales that we observe. Rotation is also hinted at in the moment 1 map presented in figure 2 of ref. 18. Although these observations probe the gas on larger scales (where most gas in the northern outflow lobe is blueshifted), it is obvious that velocities are redshifted with respect to the observer in the northwestern cavity wall and at small distances from the protostar.

The launching radius. The launching radius (footpoint radius, r_0) of the outflow is determined using a first-order polynomial fitting ($z = A(r - r_0)$; where z is the distance above the disk midplane, r is the distance from the central outflow axis and r_0 and A are free parameters) of flux-weighted positions for each channel, and in each pixel where the signal-to-noise ratio is larger than 3 (Fig. 2). Since we are primarily interested in the outflowing gas traced by ^{12}CO , we exclude the following regions of parameter space. First, we consider only gas at a velocity of more than 2 km s^{-1} offset from the systemic velocity. This is the Keplerian velocity of the disk at 100 AU, and any envelope emission on larger scales will have a velocity lower than this value. Second, we exclude velocities $> 5.5 \text{ km s}^{-1}$ with respect to the systemic velocity, since the outflow emission at these higher velocities is confined to within 20 AU of the disk midplane. This analysis does not take into account that, at any given velocity, the launching region can be extended. However, it provides a straightforward and direct estimate of where the outflow is launched. We consider only the gas located to the east of the central outflow axis, since this is the only region where we attain a sufficient signal-to-noise ratio to perform a quantitative analysis. This is also the cavity wall where the emission is most extended. The slope of the line and the crossing point (footpoint radius) of the disk midplane are thus determined by the flux density distribution across the map. We include all Nyquist sampled data points on the northern side of the disk outlined by the continuum. However, the velocity observed close to the disk surface can be much lower than the local escape velocity and we therefore also performed the analysis excluding all data points within 20 AU of the disk surface (deprojected distance). The resulting values for r_0 , however, are indifferent to the choice of cut-off height, and our conclusions are therefore not affected.

We fitted the outflowing gas by a first-order polynomial to avoid making too many assumptions about the geometrical structure of the flow. We do acknowledge that, theoretically, the gas will not follow straight lines in the immediate vicinity of the launching region. To test the robustness of our results, we also fitted a second-order polynomial to the emission ($z = A(r - r_0)^2$). The derived footpoint radii are similar and we conclude that the choice of exponent does not affect our scientific conclusions. If we instead assume an intercept of zero during the fitting procedure (that is, $z = Ar^2$; ref. 38), the goodness of fit decreases because the observed geometry is considerably steeper than can be modelled by any simple polynomial with a zero intercept (such as $z = Ar^B$).

In an independent analysis, we estimate the footpoint radius for each detected pixel in the ^{12}CO map, using equation (4) of ref. 22. The characteristic velocity in each position is taken from the computed velocity field (Fig. 1) and in each position the velocity is decomposed into two components accounting for inclination and the systemic velocity: the rotational velocity and the outflow velocity. We note that the estimated footpoint radius can be affected by entrained gas and/or asymmetries in the flow, since this increases the uncertainty on the magnitude of the velocity components. Further, such an analysis can only be carried out in the ballistic regime,

and for that reason, we mask out all pixels where the local escape velocity exceeds the outflow velocity. The estimated footpoint radius for each position is presented in Fig. 3. An illustration of where the outflow is launched is shown in Extended Data Fig. 5. In both figures, straight lines point towards the launching point.

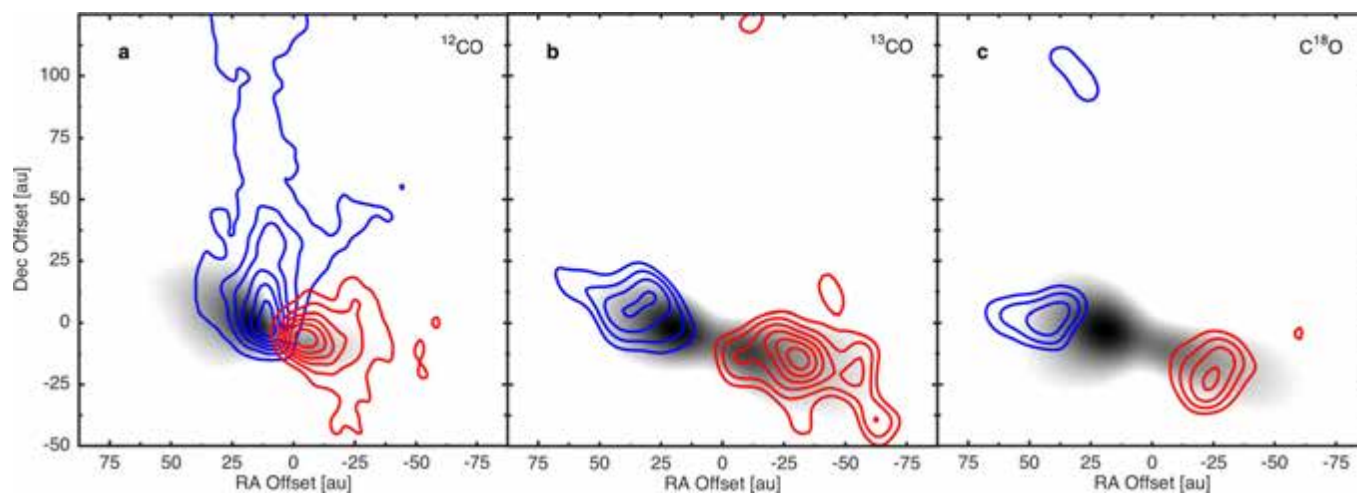
Dust continuum emission. To examine the continuum emission from the disk, we fitted a Gaussian profile to the emission as a function of the radius, deprojected by the inclination angle of the system, for all azimuthal directions. This reveals an enhancement in the emission around 20 AU at the 1 mJy per beam level (see Extended Data Fig. 3), which is consistent with the estimated footpoint radius for the lowest-velocity channels. Since the emission cannot easily be explained by an analytical function, we exclude all data points from the Gaussian fit where the enhancement is most prominent, that is, between 12 AU and 33 AU. The variation with azimuthal angle of the distance to the peak position of this enhancement is smaller than the resolution element in these observations (see Extended Data Fig. 3).

Angular momentum of the outflowing gas. To estimate the specific angular momentum of the outflowing gas, we use the ^{12}CO velocity field. The specific angular momentum is calculated as the product of the rotational velocity, and the distance to the central axis of the blueshifted outflow (see Extended Data Fig. 4). The uncertainty on the rotation velocity is dominated by the uncertainty on the inclination angle (about 10°), since the uncertainty in observed velocity is negligible in comparison.

Code availability. The code RADMC-3D, used for the Keplerian disk modelling, is available at: <http://www.ita.uni-heidelberg.de/~dullemond/software/radmc-3d/>. We have opted not to make the molecular excitation code available owing to the lack of documentation and the non-trivial nature of its usage.

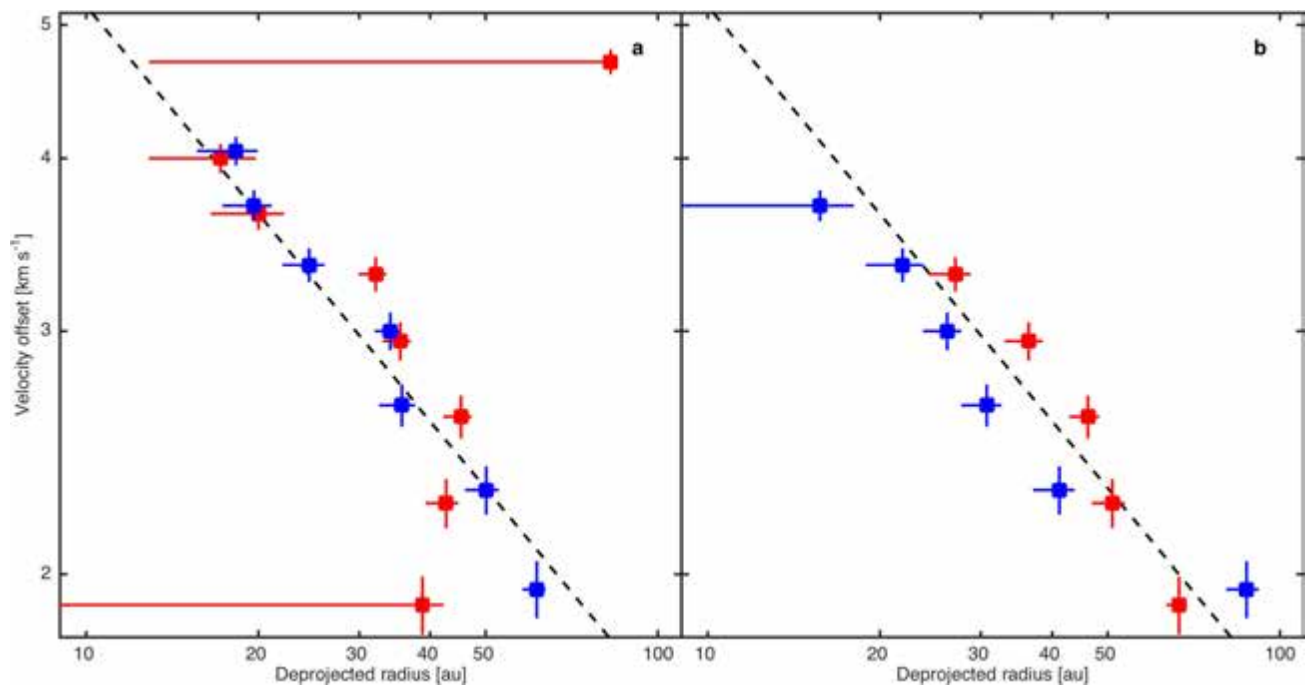
Data availability. This paper makes use of the following ALMA data: ADS/JAO.ALMA#2015.1.01415.S. The datasets generated and/or analysed during the current study are available in the ALMA archive (http://almascience.eso.org/aq/?project_code=2015.1.01415.S) and are also available from the corresponding author upon reasonable request.

31. McMullin, J. P., Waters, B., Schiebel, D., Young, W. & Golap, K. in *Astronomical Data Analysis Software and Systems* Vol. 376 *Astronomical Society of the Pacific Conference Series* (eds Shaw, R. A., Hill, F. & Bell, D. J.) 127 (2007).
32. Högbom, J. A. Aperture synthesis with a non-regular distribution of interferometer baselines. *Astrophys. J. Suppl. Ser.* **15**, 417–426 (1974).
33. San José-García, I. *et al.* Herschel-HIFI observations of high-J CO and isotopologues in star-forming regions: from low to high mass. *Astron. Astrophys.* **553**, 125–153 (2013).
34. Harsono, D., van Dishoeck, E. F., Bruderer, S., Li, Z. Y. & Jørgensen, J. K. Testing protostellar disk formation models with ALMA observations. *Astron. Astrophys.* **577**, 22–37 (2015).
35. Dullemond, C. P. & Dominik, C. Flaring vs. self-shadowed disks: the SEDs of Herbig Ae/Be stars. *Astron. Astrophys.* **417**, 159–168 (2004).
36. Bruderer, S., van Dishoeck, E. F., Doty, S. D. & Herczeg, G. J. The warm gas atmosphere of the HD 100546 disk seen by Herschel. Evidence of a gas-rich, carbon-poor atmosphere? *Astron. Astrophys.* **541**, A91 (2012).
37. Sakai, N. *et al.* Subarcsecond analysis of the infalling-rotating envelope around the class I protostar IRAS 04365+2535. *Astrophys. J.* **820**, L34 (2016).
38. Lee, C. F., Mundy, L. G., Reipurth, B., Ostriker, E. C. & Stone, J. M. CO outflows from young stars: confronting the jet and wind models. *Astrophys. J.* **542**, 925–945 (2000).



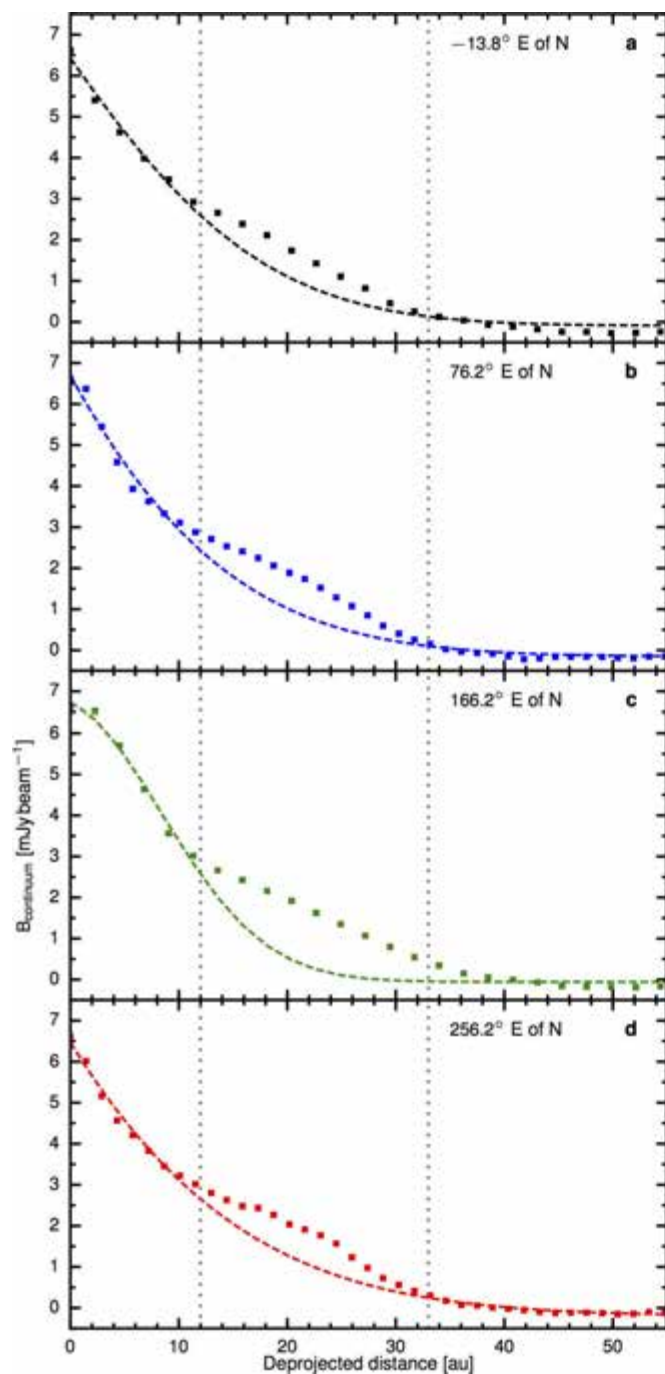
Extended Data Figure 1 | Comparison of integrated emission for ^{12}CO , ^{13}CO and C^{18}O . Contours are from 3σ in steps of 3σ for ^{12}CO (a) and 1σ for ^{13}CO (b) and C^{18}O (c). $\sigma = 4$ mJy per beam for ^{12}CO and $\sigma = 5$ mJy per beam for ^{13}CO and C^{18}O . Redshifted (red) and blueshifted (blue)

emission is integrated from 2.5 km s^{-1} to 10 km s^{-1} with respect to the systemic velocity. The corresponding integrated emission from the power-law disk model is shown in greyscale. RA, right ascension; Dec, declination.

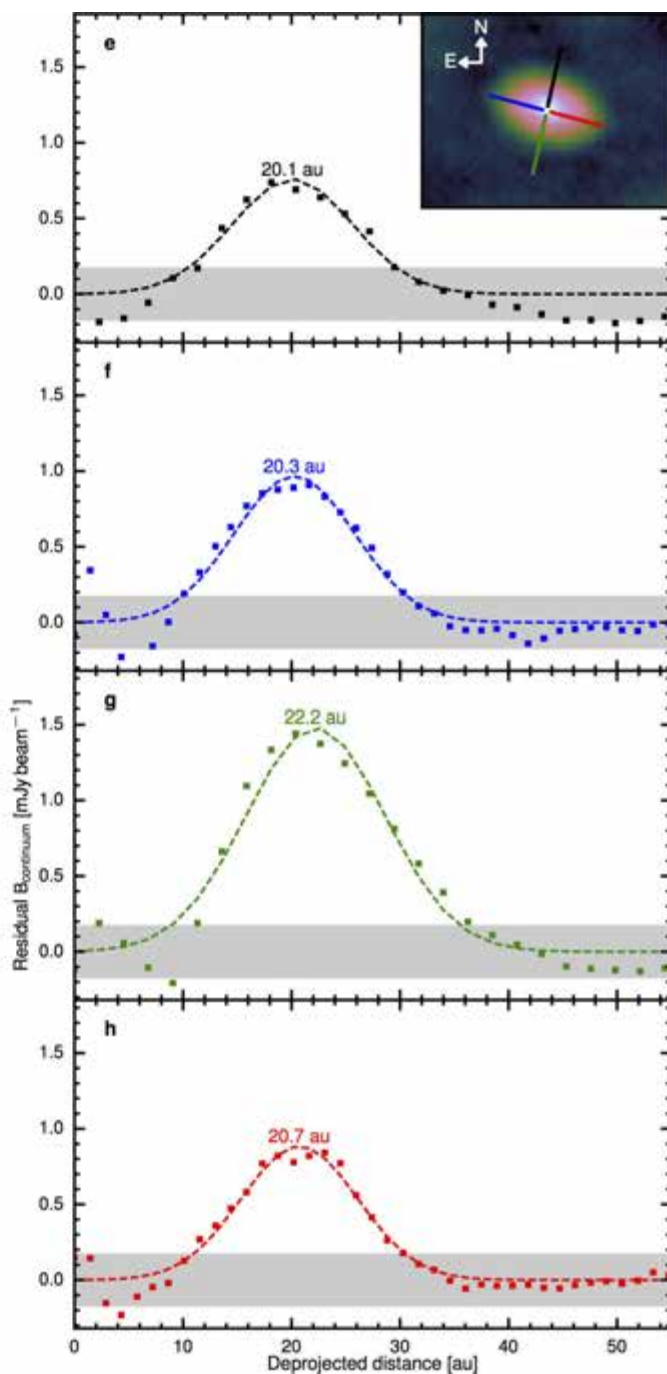


Extended Data Figure 2 | Position-velocity diagram for ¹³CO and C¹⁸O. Velocity of ¹³CO (a), and C¹⁸O (b) versus position, using an inclination angle of 55°. The dashed curve is indicative of Keplerian rotation around a 0.4M_{sun} star. The red and blue colours indicate the redshifted and

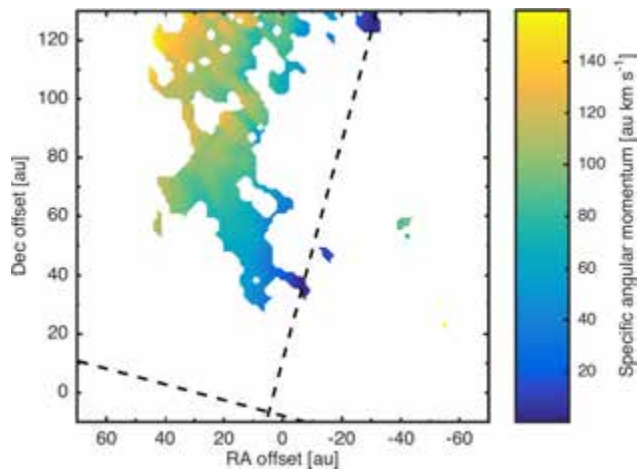
blueshifted components, respectively. Error bars show the standard deviations of the Gaussian fits in position and the velocity resolution. au, astronomical units.



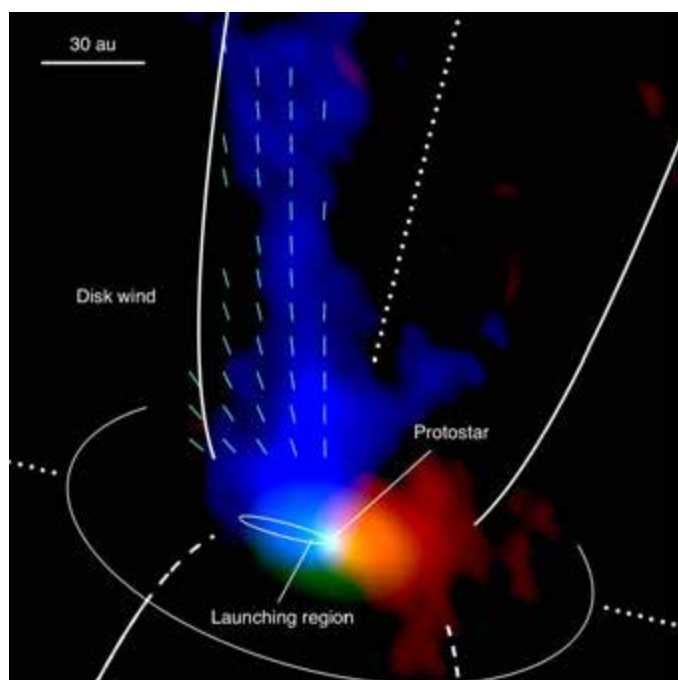
Extended Data Figure 3 | Enhancement in dust continuum emission. **a–d**, Observed radial continuum brightness profile (square data points) at the four position angles indicated in the inset at top right. ‘Position’ angle 76° corresponds to the long axis of the disk on the northeastern side where the blueshifted northern outflow is launched. A Gaussian fit is overlaid as



a dashed line. **e–h**, Residual intensity after subtracting the fits shown in the left column (square points), and a Gaussian fit (dashed line) to determine the peak location of the enhancement. The grey-filled area denotes the 2σ root-mean-square noise in the continuum map.



Extended Data Figure 4 | Specific angular momentum derived from the velocity field. The colour map shows the specific angular momentum and black dashed lines show the position angle of the outflow and the disk.



Extended Data Figure 5 | Inferred launching region of the disk wind.

This illustrative figure is overlaid on a three-colour background image, showing the blueshifted (blue) and redshifted (red) ^{12}CO emission together with the continuum emission (green). The outflow emission is integrated from $\pm(2.5\text{--}10)\text{ km s}^{-1}$ with respect to the systemic velocity 6.4 km s^{-1} . The outlines of the disk and the outflow and the axes of the disk and the outflow are indicated with white lines. Dashed blue lines are the same as in Fig. 3.

Extensive degeneracy, Coulomb phase and magnetic monopoles in artificial square ice

Yann Perrin^{1,2}, Benjamin Canals^{1,2} & Nicolas Rougemaille^{1,2}

Artificial spin-ice systems are lithographically patterned arrangements of interacting magnetic nanostructures that were introduced as way of investigating the effects of geometric frustration in a controlled manner^{1–4}. This approach has enabled unconventional states of matter to be visualized directly in real space^{5–18}, and has triggered research at the frontier between nanomagnetism, statistical thermodynamics and condensed matter physics. Despite efforts to create an artificial realization of the square-ice model—a two-dimensional geometrically frustrated spin-ice system defined on a square lattice—no simple geometry based on arrays of nanomagnets has successfully captured the macroscopically degenerate ground-state manifold of the model¹⁹. Instead, square lattices of nanomagnets are characterized by a magnetically ordered ground state that consists of local loop configurations with alternating chirality^{1,20–26}. Here we show that all of the characteristics of the square-ice model are observed in an artificial square-ice system that consists of two sublattices of nanomagnets that are vertically separated by a small distance. The spin configurations we image after demagnetizing our arrays reveal unambiguous signatures of a Coulomb phase and algebraic spin-spin correlations, which are characterized by the presence of ‘pinch’ points in the associated magnetic structure factor. Local excitations—the classical analogues of magnetic monopoles²⁷—are free to evolve in an extensively degenerate, divergence-free vacuum. We thus provide a protocol that could be used to investigate collective magnetic phenomena, including Coulomb phases²⁸ and the physics of ice-like materials.

To recover the true degeneracy associated with the square-ice model, we fabricated a series of artificial square-ice systems inspired by a previous theoretical proposition²⁹. The main idea behind that proposition is to reduce the coupling strength between perpendicularly oriented nanomagnets (J_1) while keeping the coupling strength between collinear nanomagnets (J_2) unchanged by vertically shifting one of the two sublattices of the square array (Fig. 1a). Such a height offset h makes it possible to finely tune the J_1/J_2 ratio. If $h=0$, then the system is a conventional artificial square-ice system, characterized by $J_1 > J_2$ and a magnetically ordered ground state (Fig. 1b). If h is continuously increased, then J_1 can become infinitesimally small compared to J_2 until a situation is reached where the horizontal and vertical lines of the square lattice are magnetically decoupled ($J_1=0$; Fig. 1c). Therefore, there is necessarily a critical height offset h_c at which the two coupling coefficients J_1 and J_2 are equal (Fig. 1d). On the basis of a dumbbell description of the nanomagnets, it was found²⁹ that $h_c = 0.207a$ for $l/a = 0.7$, where l is the length of the nanomagnets and a is the lattice parameter. A critical height offset of $h_c = 0.27a$ was calculated³⁰ by incorporating dipolar interactions over the entire volume of uniformly magnetized nanomagnets. However, both of these approaches neglected key experimental ingredients: the geometric properties and the micromagnetic nature of the nanomagnets were not taken into account. Here, we determine h_c from a set of micromagnetic simulations that describe the real shape and internal micromagnetic configuration of the

nanomagnets used experimentally (Methods). The main result of our calculation is that h_c strongly depends on the gap left between neighbouring magnetic elements, and is qualitatively similar to the estimate deduced from the dumbbell description (Fig. 1e and Extended Data Fig. 1), but quantitatively very different.

To recover the degeneracy of the square-ice model, we need to lithographically pattern arrays of nanomagnets in which the third dimension now plays a key part, extending artificial spin-ice systems from two to three dimensions. This additional dimension makes the fabrication and imaging of spin-ice architectures much more challenging. Our shifted artificial square-ice systems were fabricated using a two-step electron-beam lithography process (Methods and Extended Data Fig. 2). The first step is dedicated to the design of non-magnetic bases that are used to lift one sublattice of nanomagnets. The thickness of the bases determines the final height offset h (the base thicknesses used here are 60 nm, 80 nm and 100 nm). The second step

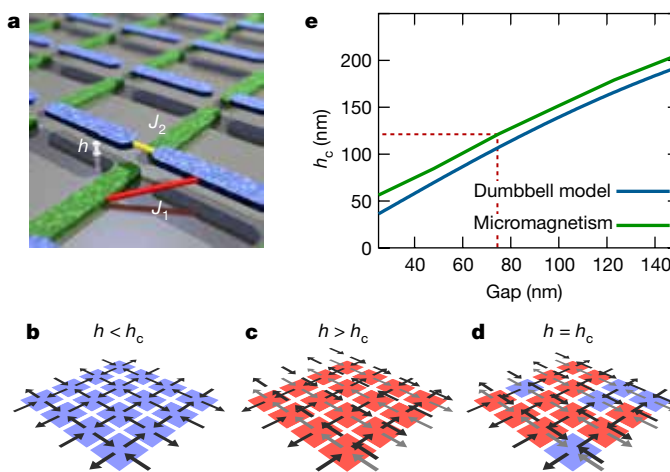


Figure 1 | Role of the nearest-neighbour coupling strength. **a**, Schematic of artificial square ice (inspired by ref. 29) in which one of the two sublattices (blue) is shifted vertically by a height offset h above the other (green). The nearest-neighbour coupling strengths between orthogonal (J_1) and collinear (J_2) nanomagnets are indicated in red and yellow, respectively. **b–d**, Ground states of the models associated with the conditions $J_1 > J_2$ ($h < h_c$; **b**), $J_1 < J_2$ ($h > h_c$; **c**) and $J_1 = J_2$ ($h = h_c$; **d**). Blue and red squares correspond to type-I and -II vertices, respectively. Black arrows indicate the local spin directions. Grey arrows represent the projection of the spin directions of the shifted sublattice on the $h=0$ plane. **e**, Plot showing the critical height offset h_c that is required to recover ice-like physics, as a function of the gap separating nearest-neighbour nanomagnets. Results derived from micromagnetic simulations (green) and from a dumbbell description (blue) of the nanomagnets are compared. The red dashed line indicates the value of h_c that we expect to observe in our experiments, given that the gap between nearest-neighbour nanomagnets is 75 nm. The length, width and thickness of our nanomagnets are 500 nm, 100 nm and 30 nm, respectively. More details are provided in the Methods.

consists of depositing the nanomagnets on a square lattice in such a way that one sublattice is grown atop the non-magnetic bases and the other is grown on the substrate. On each sample, a reference square lattice with $h = 0$ is patterned for direct comparison with the shifted arrays. Magnetic images were obtained using magnetic force microscopy (Fig. 2a) after demagnetizing the arrays in an in-plane oscillating magnetic field with slowly decaying amplitude (Methods). All of the arrays were demagnetized simultaneously to ensure identical field history between samples.

The three shifted arrays we studied were demagnetized four times to improve the statistics and to test the reproducibility of the experimental observations, and we systematically imaged the reference array ($h = 0$) present on each sample to check the efficiency of the field demagnetization protocol. For these 12 realizations, the reference arrays were always found in a magnetic configuration close to the ordered antiferromagnetic ground state (Fig. 1b). A typical magnetic state is shown in Fig. 2b, in which a domain boundary separating two anti-phase domains is observed. Consequently, type-I vertices are present everywhere, except in the domain wall formed by type-II vertices (for definitions of type-I and -II vertices, see the inset of Fig. 1a). Our demagnetization protocol is therefore efficient and brings the system into a low-energy manifold, with large patches of the ground-state configuration, similar to what is found in thermally active artificial spin ices^{10,21,24–26}. For the reference arrays, the density of type-I, -II, -III and -IV vertices are, on average, 86%, 12.5%, 1.5% and 0%, respectively (Fig. 2c; type-III (-IV) vertices refer to vertices with three (four) in or three (four) out spin configurations¹), and the mean size of type-I domains is about 87 vertices. Consequently, the residual magnetization is low, typically 3% in both the vertical and horizontal directions. The computed magnetic structure factor (Methods and Extended Data Fig. 3), averaged over the 12 different reference arrays, shows clear magnetic Bragg peaks located at the corners of the Brillouin zone (Fig. 3a).

Figure 2c shows the variation in vertex density ρ when h is increased, revealing a clear trend: the density of type-I vertices continuously decreases whereas the density of type-II vertices increases. For $h = 60$ nm, the physics is essentially unchanged from the $h = 0$ case: type-I vertices are the most prevalent and form patches of the antiferromagnetic ground state, although the average size of the ordered domains decreases to 15 vertices—6 times smaller than for the reference arrays. The corresponding magnetic structure factor shows a spreading of the magnetic Bragg peaks associated with antiferromagnetic ordering, but the peaks remain located at the corners of the Brillouin zone (Fig. 3b). This result is consistent with the predictions from micromagnetic simulations (Fig. 1e), which indicate that the ground state is expected to be the antiferromagnetic ordered configuration when $h = 60$ nm.

For $h = 80$ nm, the population of type-II vertices (52%) becomes higher than that of type-I vertices (39%); type-II patches start to form and the spatial extent of type-I domains is further reduced. The magnetic Bragg peaks in the magnetic structure factor have almost disappeared, which is an indication that the spin configurations have started to be disordered. If the background intensity in the magnetic structure factor becomes more diffuse, then it develops a structure with geometric features that resemble those expected from the square-ice model (Fig. 3c, e and Methods). This result is consistent with the micromagnetic simulations: although the ground state is expected to be the antiferromagnetic ordered configuration, the magnetic configuration is disordered after demagnetizing the array, as h approaches h_c (J_1 starts to compare with J_2).

The similarity to the square-ice model becomes more evident for $h = 100$ nm (Fig. 3d, e). Contrary to all previous results, which demonstrate that square lattices of nanomagnets are magnetically ordered in their low-energy manifold, we show that our artificial square ice is highly disordered. The magnetic Bragg peaks in the magnetic structure factor have totally disappeared for $h = 100$ nm and the diffuse

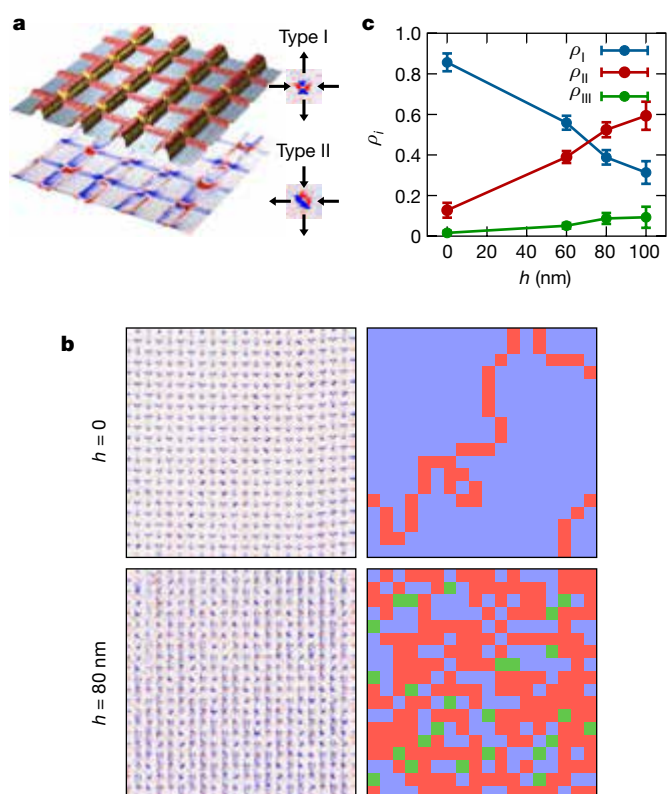


Figure 2 | Experimental results. **a**, Topography (atomic force microscopy; top) and magnetic (magnetic force microscopy; bottom) images of our artificial realization of the square-ice model. In the topography image, the nanomagnets appear red, the bases are yellow and the substrate is grey. In the magnetic image, the magnetic contrast appears in blue and red for negative and positive magnetic charges, respectively. Typical contrasts obtained on type-I and -II vertices are shown in the inset. Type-I (-II) vertices correspond to local two-in/two-out spin configurations carrying zero (net) magnetic moment. **b**, Magnetic images (raw data on the left and the corresponding analysis on the right) for a height offset of $h = 0$ (top) and $h = 80$ nm (bottom). For $h = 0$, most of the vertices are type-I (blue), and a domain boundary separating anti-phase domains is clearly visible (type-II vertices are shown in red). For $h = 80$ nm, the magnetic state appears disordered; type-III vertices are coloured in green. **c**, Analysis of the vertex density of type- i vertices ρ_i as a function of the height offset h . The points represent the mean and the error bars represent the standard deviation calculated from the four demagnetizations.

background is strongly structured. However, on the basis of the micromagnetic simulations presented above, the ground state of artificial square ice with $h = 100$ nm is expected to be ordered. We interpret this difference between observation and prediction as a consequence of the kinetics associated with the spin dynamics when the sample is demagnetized under a rotating magnetic field. During the demagnetization protocol, spins are reversed via an avalanche process that favours the formation of straight lines (Methods and Supplementary Videos 1 and 2). Type-II vertices are then stabilized by the external magnetic field at the expense of type-I vertices, even though type-I vertices have a slightly lower energy. In other words, our protocol shifts the critical value h_c , at which the transition to the disordered phase is expected.

Our square ice exhibits all of the characteristics of a dipolar algebraic spin liquid. In such a spin liquid, there are locations in reciprocal space where the magnetic structure factor behaves non-analytically and has the shape of a pinch point. These pinch points are visible in our experimental map (Fig. 3d, circle) and are indicative of a Coulomb phase²⁸. To gain more insight into the observed physics, we quantitatively compare, on the experimental and theoretical maps, q -scans of the magnetic structure factor along two orthogonal lines passing through these pinch

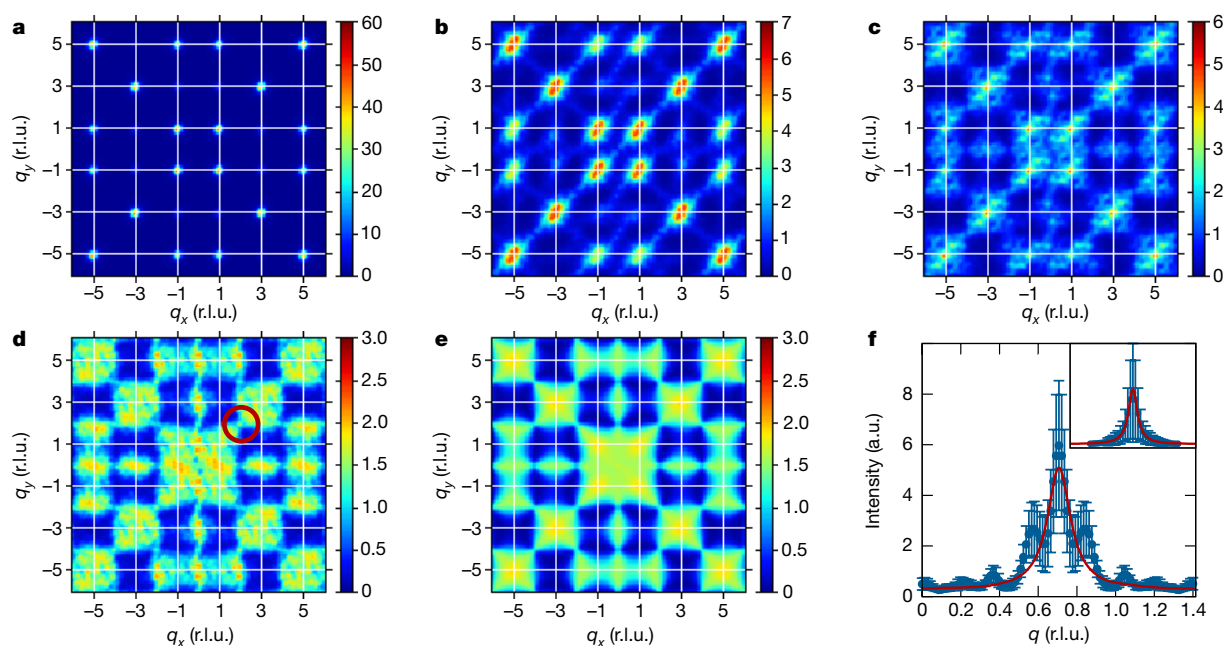


Figure 3 | Magnetic structure factors and pinch-point analysis.

a–d, Magnetic structure factors deduced from the experimental images for $h = 0$ (**a**), $h = 60$ nm (**b**), $h = 80$ nm (**c**) and $h = 100$ nm (**d**). The colour scale refers to the intensity at a given point (q_x, q_y) of reciprocal space. **e**, Computed magnetic structure factor averaged over 1,000 random, decorrelated spin configurations that satisfy the ice rule (Methods). **f**, Experimental (main plot) and theoretical (inset) intensity profiles

across the pinch point highlighted by a red circle in **d**. The points represent the mean and the error bars represent the standard deviation calculated from the four demagnetizations for the experimental data and the 1,000 random ice-rule configurations for the theoretical data. The red curves are single-peaked Lorentzian fits of the pinch points. The q axis corresponds to a scan from $(3/2, 5/2)$ to $(5/2, 3/2)$ in reciprocal space. r.l.u., reciprocal lattice unit; a.u., arbitrary units.

points (Methods, Extended Data Figs 4 and 5 and Extended Data Table 1). Whereas the theoretical scan reveals a sharp peak associated with a correlation length of $\xi_{\text{theo}} = 5.2a \pm 4\%$, the experimental scan displays a broader peak associated with a shorter correlation length of $\xi_{\text{exp}} = 4.4a \pm 12\%$. The broader peaks indicate the presence of local excitations, that is, a finite density of classical monopoles within a Coulomb phase.

Experimentally, we observe a similar density of $+2$ and -2 monopoles for all of the arrays (although they do not systematically obey charge neutrality owing to their finite size), with the density increasing as we retrieve the degeneracy of the square-ice model (see h dependence of type-III vertices in Fig. 2c). The variation in the monopole density we measure is not random and appears to be robust when comparing successive demagnetization protocols. The

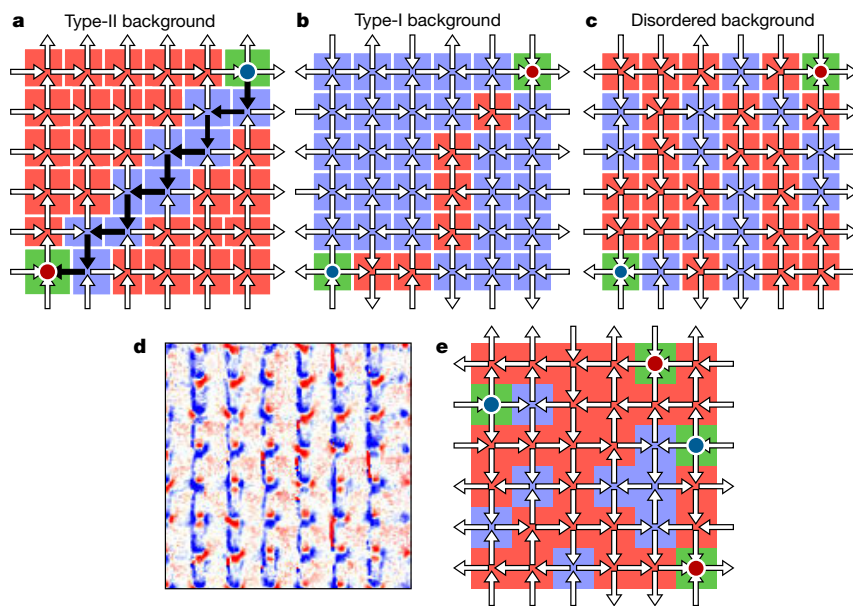


Figure 4 | Magnetic monopoles in square-ice systems. **a–c**, Monopole/anti-monopole pair (red and blue circles) in a magnetically saturated background (type-II background; **a**), in the antiferromagnetic ground state (type-I background; **b**) and within a disordered manifold (disordered background; **c**). Blue, red and green squares indicate type-I, -II and -III

vertices, respectively. The black arrows in **a** illustrate a chain of reversed spins. **d**, Experimental spin configurations for $h = 100$ nm showing two pairs of oppositely charged monopoles. **e**, Analysis of the configuration in **d**. Monopoles appear as red and blue circles on top of a green square.

densities are fairly high when approaching the spin-liquid phase (Fig. 2c), meaning that we do not bring the system into its massively degenerate ground-state manifold. Instead, the imaged spin configurations are characteristic of excited states embedded within a Coulomb phase. The monopoles that we observe in our arrays differ substantially from those that have previously been visualized in artificial square ices^{1,20–26}. All of the monopoles reported so far are high-energy local configurations evolving in an uncharged, but magnetically ordered, vacuum (Fig. 4a, b), characterized by a magnetic structure factor that contains only magnetic Bragg peaks (Methods).

Our system has very distinct behaviour: the monopoles we observe are free to move into a spin-liquid state, that is, a massively degenerate, disordered low-energy manifold (Fig. 4c). They are therefore particle-like objects present in a diffuse but structured magnetic structure factor, free of any Bragg peaks (Fig. 3d). An example experimental configuration is shown in Fig. 4d and is schematized in Fig. 4e. Two pairs of oppositely charged monopoles are present in this disordered magnetic configuration containing type-I and -II vertices. It is not possible to determine the path that these monopoles have followed during the demagnetization process, because the trace of reversed spins (often referred to as a Dirac string) has been erased by the magnetic disorder. This is in contrast to the aforementioned cases for which monopoles evolve within an ordered spin configuration and for which the influence of the field or temperature can be unambiguously visualized (Fig. 4a, b). Here, there is no way of knowing the trajectory of the magnetic monopoles and, consequently, it is not even possible to pair two oppositely charged monopoles.

This result raises interesting, crucial questions. We envision that, if similar artificial, shifted square-ice systems could be made thermally active, the dynamics of these de-confined, interacting quasi-particles could be investigated in real space and time. We then wonder whether a typical distance between oppositely charged monopoles would be established at thermodynamic equilibrium and whether this distance could be linked to the correlation length that was deduced from the analysis of the width of the pinch point (Extended Data Fig. 6 and Methods). It would also be interesting to study how these quasi-particles nucleate, propagate and annihilate with their anti-particle, and to directly observe how their interactions affect the disordered background.

We have shown that shifted magnetic square lattices offer the possibility to tune the nearest-neighbour coupling strength and, in particular, to experimentally realize the seminal square-ice model. The fabrication of thermally active, shifted square ice in the future would enable the thermodynamics and dynamics of the low-energy manifolds and the recombination of their topological excitations to be investigated (Methods). Finally, our work demonstrates that artificial loop models, such as the square-ice model, are not beyond reach, thanks to lithography engineering. These models have numerous extensions in very different fields of research, including polymer physics, topological quantum computing, self-avoiding random walks and Schramm–Loewner evolution. Implementing loop models is therefore of broad interest in physics and chemistry, and our contribution illustrates that magnetic versions of these loop models are experimentally accessible.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 July; accepted 12 October 2016.

Published online 28 November 2016.

- Wang, R. F. *et al.* Artificial spin ice in a geometrically frustrated lattice of nanoscale ferromagnetic islands. *Nature* **439**, 303–306 (2006).
- Nisoli, C., Moessner, R. & Schiffer, P. Artificial spin ice: designing and imaging magnetic frustration. *Rev. Mod. Phys.* **85**, 1473–1490 (2013).
- Heyderman, L. J. & Stamps, R. L. Artificial ferroic systems: novel functionality from structure, interactions and dynamics. *J. Phys. Condens. Matter* **25**, 363201 (2013).

- Cummings, J., Heyderman, L. J., Marrows, C. H. & Stamps, R. L. Focus on artificial frustrated systems. *New J. Phys.* **16**, 075016 (2014).
- Ladak, S., Read, D. E., Perkins, G. K., Cohen, L. F. & Branford, W. R. Direct observation of magnetic monopole defects in an artificial spin-ice system. *Nat. Phys.* **6**, 359–363 (2010).
- Mengotti, E. *et al.* Real-space observation of emergent magnetic monopoles and associated Dirac strings in artificial kagome spin ice. *Nat. Phys.* **7**, 68–74 (2011).
- Rougemaille, N. *et al.* Artificial kagome arrays of nanomagnets: a frozen dipolar spin ice. *Phys. Rev. Lett.* **106**, 057209 (2011).
- Zhang, S. *et al.* Crystallites of magnetic charges in artificial spin ice. *Nature* **500**, 553–557 (2013).
- Montaigne, F. *et al.* Size distribution of magnetic charge domains in thermally activated but out-of-equilibrium artificial spin ice. *Sci. Rep.* **4**, 5702 (2014).
- Drisko, J., Daunheimer, S. & Cummings, J. FePd₃ as a material for studying thermally active artificial spin ice systems. *Phys. Rev. B* **91**, 224406 (2015).
- Anghinolfi, L. *et al.* Thermodynamic phase transitions in a frustrated magnetic metamaterial. *Nat. Commun.* **6**, 8278 (2015).
- Chioar, I. A. *et al.* Kinetic pathways to the magnetic charge crystal in artificial dipolar spin ice. *Phys. Rev. B* **90**, 220407(R) (2014).
- Zhang, S. *et al.* Perpendicular magnetization and generic realization of the Ising model in artificial spin ice. *Phys. Rev. Lett.* **109**, 087201 (2012).
- Chioar, I. A. *et al.* Nonuniversality of artificial frustrated spin systems. *Phys. Rev. B* **90**, 064411 (2014).
- Chioar, I. A., Rougemaille, N. & Canals, B. Ground-state candidate for the dipolar kagome Ising antiferromagnet. *Phys. Rev. B* **93**, 214410 (2016).
- Gilbert, I. *et al.* Emergent ice rule and magnetic charge screening from vertex frustration in artificial spin ice. *Nat. Phys.* **10**, 670–675 (2014).
- Brooks-Bartlett, M. E., Banks, S. T., Jaubert, L. D. C., Harman-Clarke, A. & Holdsworth, P. C. W. Magnetic-moment fragmentation and monopole crystallization. *Phys. Rev. X* **4**, 011007 (2014).
- Canals, B. *et al.* Fragmentation of magnetism in artificial kagome dipolar spin ice. *Nat. Commun.* **7**, 11446 (2016).
- Lieb, E. H. Residual entropy of square ice. *Phys. Rev.* **162**, 162–172 (1967).
- Nisoli, C. *et al.* Effective temperature in an interacting vertex system: theory and experiment on artificial spin ice. *Phys. Rev. Lett.* **105**, 047205 (2010).
- Morgan, J. P., Stein, A., Langridge, S. & Marrows, C. H. Thermal ground-state ordering and elementary excitations in artificial magnetic square ice. *Nat. Phys.* **7**, 75–79 (2011).
- Budrikis, Z., Politi, P. & Stamps, R. L. Diversity enabling equilibration: disorder and the ground state in artificial spin ice. *Phys. Rev. Lett.* **107**, 217204 (2011).
- Budrikis, Z. *et al.* Domain dynamics and fluctuations in artificial square ice at finite temperatures. *New J. Phys.* **14**, 035014 (2012).
- Farhan, A. *et al.* Direct observation of thermal relaxation in artificial spin ice. *Phys. Rev. Lett.* **111**, 057204 (2013).
- Porro, J. M., Bedoya-Pinto, A., Berger, A. & Vavassori, P. Exploring thermally induced states in square artificial spin-ice arrays. *New J. Phys.* **15**, 055012 (2013).
- Kapakliis, V. *et al.* Thermal fluctuations in artificial spin ice. *Nat. Nanotechnol.* **9**, 514–519 (2014).
- Castelnovo, C., Moessner, R. & Sondhi, S. L. Magnetic monopoles in spin ice. *Nature* **451**, 42–45 (2008).
- Henley, C. L. The “Coulomb phase” in frustrated systems. *Annu. Rev. Condens. Matter Phys.* **1**, 179–210 (2010).
- Möller, G. & Moessner, R. Artificial square ice and related dipolar nanoarrays. *Phys. Rev. Lett.* **96**, 237202 (2006).
- Thonig, D., Reißaus, S., Mertig, I. & Henk, J. Thermal string excitations in artificial spin-ice square dipolar arrays. *J. Phys. Condens. Matter* **26**, 266006 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Agence Nationale de la Recherche through project number ANR12-BS04-009 ‘Frustrated’. We acknowledge support from the Nanofab team at the Institut NÉEL and thank S. Le-Denmat and O. Fruchart for technical help during atomic force microscope and magnetic force microscope measurements.

Author Contributions B.C. and N.R. conceived the project. Y.P. was in charge of the sample fabrication and characterization, the magnetic imaging measurements and the analysis of the data. All authors contributed to the preparation of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.R. (nicolas.rougemaille@neel.cnrs.fr).

Reviewer Information *Nature* thanks C. Nisoli, A. Ramirez and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Micromagnetic simulations. Micromagnetic effects have been shown to play an important part in artificial spin ices, for example by modifying the coupling strength between neighbouring elements⁷, by inducing chirality³¹ and by controlling magnetization reversal processes³². Our micromagnetic simulations are based on a finite-difference approach whereby the system is discretized into rectangular cells. We used the three-dimensional solver of the OOMMF free software from the National Institute for Standards and Technology (NIST)³³. We computed the four energy levels E_i ($i \in \{1, 2, 3, 4\}$) of an isolated square vertex composed of four permalloy nanomagnets in magnetostatic interaction. The nanomagnets have dimensions of $500 \text{ nm} \times 100 \text{ nm} \times 30 \text{ nm}$. The gap g between the nanomagnets is defined as the distance between the extremities of the nanomagnet and the centre of the vertex. In all calculations, the exchange stiffness is set to 10 pJ m^{-1} , the magnetocrystalline anisotropy is neglected, spontaneous magnetization M_s is $8 \times 10^5 \text{ A m}^{-1}$ (1.0 T) and the damping coefficient is set to 1 to speed up convergence. Given that the volume of a nanomagnet is $1.44 \times 10^{-21} \text{ m}^3$, our nanomagnets carry a magnetic moment of about $\mu = 11.5 \times 10^{-16} \text{ A m}^2$. To limit the influence of numerical roughness, the mesh size was reduced to $3 \text{ nm} \times 3 \text{ nm} \times 15 \text{ nm}$. No qualitative difference was observed when reducing the mesh size in the z direction.

Dumbbell model. In Fig. 1e, we plot the value of the critical height h_c as a function of the gap between neighbouring nanomagnets in two cases: within a full micromagnetic approach and using a dumbbell description. For the dumbbell description, we followed the procedure reported in ref. 29; we reproduced the results therein and present them in Extended Data Fig. 1, in which the ratio J_1/J_2 is shown as a function of l/a and h/a , where l is the length of the nanomagnets (that is, the distance between the two magnetic charges) and a is the lattice parameter. The condition $J_1 = J_2$ is indicated by the dark line.

Sample fabrication. The shifted artificial square ices were fabricated using a two-step electron-beam lithography process (Extended Data Fig. 2). The first step is dedicated to the design of the non-magnetic titanium/gold bases. Their shape was optimized to maximize the probability of successfully aligning the nanomagnets that were deposited in the second step. After exposing and developing a PMMA (poly(methyl methacrylate)) layer, the bases were deposited by electron-beam evaporation. To obtain a strong contrast in the scanning electron microscope through the PMMA resist, the top of the bases was made of a 50-nm-thick gold layer. The titanium thickness was then adjusted to obtain the desired height offset between the two sub-lattices. An ultrasound-assisted lift-off at 80°C in the remover revealed the bases. A new layer of PMMA resist was then spin-coated on the sample. Arrays of nanomagnets were patterned atop the base areas and deposited by electron-beam evaporation after developing the resist. The ferromagnetic layer was made of 30-nm-thick $\text{Ni}_{80}\text{Fe}_{20}$ and a 3-nm-thick aluminium capping layer was deposited to avoid oxidation. A 5-nm-thick titanium layer between the bases and the permalloy layer was used to enhance its adherence. Finally, a similar lift-off process removed the unwanted material from the samples. For the two steps, a 150-nm-thick PMMA layer was spin-coated on a Si(100) substrate. Electron-beam lithography steps were done using a Raith LEO scanning electron microscope. The two layers were aligned manually in translation and rotation using adapted cross-marks. Our success rate using this method was around 20% for the nanomagnet sizes used here. Ti, Al, Au and $\text{Ni}_{80}\text{Fe}_{20}$ were deposited with an evaporation rate of about 0.1 nm s^{-1} (pressure of 10^{-5} mbar).

Magnetic imaging. Magnetic images were obtained using a NT-MDT magnetic force microscope. Custom-made low-moment magnetic tips were used to avoid magnetization reversal in the nanomagnets while scanning the arrays. The magnetic layer of the tips was made from a 30- or 50-nm-thick CoCr alloy.

Demagnetization protocol. Prior to their imaging, the samples were demagnetized using an in-plane oscillating field (250-mHz sine function) with decaying amplitude while in rotation at a typical frequency of several tens of hertz. A very slow field ramp was used to decrease the amplitude of the applied external magnetic field from 100 mT (well above the coercive field of our nanomagnets) to 0 in 72 h. As demonstrated by the magnetic configurations obtained on the different reference samples ($h = 0$), our protocol efficiently brings the arrays within their low-energy manifold. In fact, our field protocols seem to be as efficient as the thermal annealing procedures used previously to capture the low-energy physics in square arrays of nanomagnets. It is possible that the free boundary conditions in our system, which contains several hundred nanomagnets, ease the demagnetization process by expelling high-energy configurations out of the lattice. However, several experimental observations suggest that the finite size of our arrays is not the key ingredient for the efficiency of the demagnetization protocol. First, we observe monopole defects close to the edges and in the core of the square lattice. There is no obvious sign that lattice edges ease the expulsion of monopoles. Second, contrary to most previous work, we use a very long demagnetization protocol (typically 72 h). The main reason for this is that the spin-spin correlations continue

to evolve even after several hours of demagnetization, consistent with what has been observed previously^{34,35}. It seems that, in general, allowing a large number of spin-flip events is the key ingredient to reaching low-energy manifolds in artificial spin systems. For example, to demagnetize kagome lattices, 110-h demagnetization protocols helped us to reach an effective temperature of about $0.06 J_{\text{nn}}$ (where J_{nn} is the nearest-neighbour coupling strength) and to observe spin fragmentation¹⁸. With shorter protocols, fragmentation was not observed because the associated (effective) temperature remained too high. For our square arrays, we find that three days is the optimum protocol length; we do not see substantial differences when applying longer demagnetization protocols.

Numerical demagnetization. To interpret our experimental data, we performed numerical simulations of the field demagnetization protocol. We consider a square array of magnetic point dipoles carrying a magnetic moment μ . The interaction energy E_{ij} between two spins i and j separated by a distance r_{ij} is of dipolar-type, with a cut-off radius r_c :

$$E_{ij} = \begin{cases} \frac{1}{r_{ij}^3} \left[\mu_i \cdot \mu_j - \frac{3}{r_{ij}^2} (\mu_i \cdot r_{ij})(\mu_j \cdot r_{ij}) \right] & \text{if } r_{ij} < r_c \\ 0 & \text{otherwise} \end{cases}$$

The local field H_{loc}^i felt by the dipole μ_i is the sum of all dipolar fields coming from the surrounding spins plus the external applied magnetic field. Each dipole μ_i behaves as an Ising pseudo-spin with its own switching field H_{sw}^i . Then, if

$$\frac{\mu_i}{\|\mu_i\|} \cdot H_{\text{loc}}^i > H_{\text{sw}}^i$$

the spin i is flipped. Following previous work²², we introduce a Gaussian distribution of the switching field H_{sw}^i to allow the system to approach its low-energy manifold. The probability $P(H_{\text{sw}}^n)$ for a spin n to have a switching field H_{sw}^n is

$$P(H_{\text{sw}}^n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(H_{\text{sw}}^n - H_{\text{sw}})^2}{2\sigma^2}\right]$$

where H_{sw} is the average switching field and σ the spreading (standard deviation) of the distribution. For lattices with $h = 0$ (h being the height offset), the ground state is ordered and σ provides a control on the density of nucleation sites during the demagnetization protocol. Our experimental results are well reproduced if σ is set to $0.1 H_{\text{sw}}$. In the simulations, the field ramp decreases linearly through 10^6 steps and 10^4 turns. Numerical demagnetizations showing the reversal of spin chains and the complete protocol are provided as Supplementary Videos 1 and 2. The same colour code is used here: the four vertex types (I–IV) are represented by squares coloured in blue, red, green and yellow, respectively. As in the experiments, type-IV vertices are never observed in the simulations.

Magnetic structure factor. We define the magnetic structure factor as in neutron scattering experiments, in which the spin correlations perpendicular to the diffusion vector \mathbf{q} are measured. We therefore define a perpendicular spin component $S_{i\alpha}^\perp$:

$$S_{i\alpha}^\perp = S_{i\alpha} - (\hat{\mathbf{q}} \cdot S_{i\alpha}) \hat{\mathbf{q}} \quad (1)$$

where $\hat{\mathbf{q}}$ is the unit vector along the diffusion vector \mathbf{q} :

$$\hat{\mathbf{q}} = \frac{\mathbf{q}}{\|\mathbf{q}\|}$$

Extended Data Fig. 3a shows the geometric construction of the vectors involved in equation (1). The intensity $I(\mathbf{q})$ scattered at location \mathbf{q} in reciprocal space is defined as

$$I(\mathbf{q}) = \frac{1}{N} \sum_{i,j=1}^{N/2} \sum_{\alpha,\beta=1}^2 S_{i\alpha}^\perp \cdot S_{j\beta}^\perp \exp(i\mathbf{q} \cdot \mathbf{r}_{i\alpha,j\beta}) \quad (2)$$

in which i and j scan all $N/2$ unity cells, and α and β the two sites of each cell. However, obtaining a more convenient form for equation (2) would enable a direct calculation starting from a magnetic configuration. $I(\mathbf{q})$ can be split into two parts $I = I^\parallel - I^\perp$ with

$$I^\parallel(\mathbf{q}) = \frac{1}{N} \sum_{\alpha,\beta=1}^2 g_\alpha(\mathbf{q}) g_\beta^*(\mathbf{q})$$

$$I^\perp(\mathbf{q}) = \frac{1}{N} \sum_{\alpha,\beta=1}^2 p_\alpha(\mathbf{q}) p_\beta^*(\mathbf{q})$$

where

$$g_{\alpha}(\mathbf{q}) = \sum_{i=1}^N \sigma_{i\alpha} \exp(i\mathbf{q} \cdot \mathbf{r}_{ij})$$

$$p_{\alpha}(\mathbf{q}) = \sum_{i=1}^N \hat{\mathbf{q}} \cdot \mathbf{S}_{i\alpha} \exp(i\mathbf{q} \cdot \mathbf{r}_{ij})$$

and $\sigma_{i\alpha}$ is the Ising variable (± 1) of the site ($i\alpha$). In this split form, I is a real quantity. To compute $I(\mathbf{q})$ diagrams in reciprocal space, we calculate the quantity $I = g_1^2 + g_2^2 - (p_1 + p_2)^2$ at several \mathbf{q} locations. The magnetic structure factor reported in Extended Data Fig. 3b is composed of a matrix of 120×120 points covering an area of $q_x, q_y \in [-6\pi, 6\pi]$. This area is 36 times larger than the first Brillouin zone.

Generating low-energy magnetic configurations. The magnetic structure factor of the square-ice model¹⁹ was computed by averaging 1,000 low-energy spin configurations that satisfy the ice rule everywhere. To do so, we start from a magnetically saturated configuration and then flip a number N of randomly chosen spin loops. These loop flips are necessary to ensure that all of the generated spin configurations satisfy the ice rule. Because our lattice has free boundary conditions, the procedure leads to open (crossing the array) or closed loops (Extended Data Fig. 4). Both loops are used to generate a low-energy spin configuration. To decorrelate the initial (saturated) and final spin configurations, we take N to be of the order of the number of spins present in the array (840).

Pinch points and correlation length. The magnetic structure factor shown in Extended Data Fig. 3b is averaged over 1,000 ice-rule configurations. Pinch points located at the centre Γ of the Brillouin zone are clearly visible, indicating the existence of a Coulomb phase and algebraic spin–spin correlations, that is, a correlated, disordered magnet within which spin–spin correlations decay like point-dipole interactions³⁶. The finite size of our arrays has consequences for the magnetic structure factor, in particular for the width of the pinch points. Extended Data Fig. 5 shows the influence of the lattice size L on the width of the pinch points, which narrow as the lattice size is increased. This width of the pinch points can be linked to a correlation length ξ in the system³⁷. This correlation length can be extracted from a Lorentzian fit to the intensity profile passing through a pinch point:

$$I(\mathbf{q}) = A \frac{\xi^{-2}}{(q - q_0)^2 + \xi^{-2}} + B$$

where A and B are constants, q_0 is the location of the pinch point in reciprocal space, and q is the diffusion vector. The correlation lengths deduced for different lattice sizes are reported in Extended Data Table 1. Uncertainties represent the variability within the 1,000 sampled spin states.

Long-range dipolar interactions. Although the square-ice model is a vertex model, here we have interacting dipoles with long-range effects. It is not clear whether the infinite range of the dipolar interaction affects the physics of the Coulomb phase. This question has been addressed in pyrochlore systems and is often referred to as the projective equivalence³⁸, meaning that models of dipolar spin ice and nearest-neighbour spin ice are almost equivalent down to very low temperatures. However, as far as we know, this question has been addressed only to a small extent in two-dimensional square ice³⁹. It has been shown numerically³⁹ that the Coulomb phase remains present down to low temperatures, even in the case of long-range dipolar interactions, and that the system eventually orders at very low temperatures. In our system, we envision two different magnetic orderings depending on the sign of $h - h_c$. For $h < h_c$ we expect antiferromagnetic ordering (Fig. 1b), whereas for $h > h_c$ we expect to see ferromagnetic lines in the system (Fig. 1c). But, similarly to spin-ice compounds and other artificial spin systems, reaching the ordered ground state is extremely difficult in practice, if not impossible. When we clearly observe the Coulomb phase, we have no indication of emerging magnetic order, which would indicate that our system could be described, to first order, by a short-range vertex model. This is in contrast to artificial kagome spin-ice systems in which the dipolar nature of the spin–spin interaction is clearly evidenced.

Magnetic monopoles. The monopoles we observe are very different from those described previously, from which two cases may be distinguished. In the first case, observation of monopoles in artificial square ice was achieved after saturating the arrays using a magnetic field applied along a (11)-like direction and by subsequently applying a field in the opposite direction with amplitude close to the coercive field of the system. The protocol then induced random nucleations of monopoles and triggered an avalanche process⁴⁰. This protocol leads to unidirectional motion of the monopoles, which leave behind them chains of reversed spins often referred to as Dirac strings. Similar results have been obtained in thermally active arrays that have been magnetically saturated²⁴. There, monopoles are metastable objects, created on purpose, embedded within a magnetically saturated state, that is, a spin configuration containing mainly type-II vertices (Fig. 4a). In the second case,

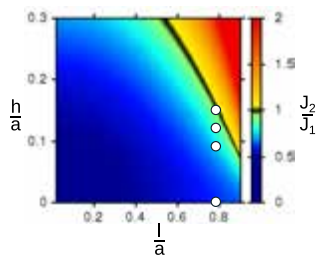
observation of monopoles in artificial square ice was achieved in arrays approaching the antiferromagnetic ordered ground state after being demagnetized or annealed²⁴. These monopoles do not necessarily move along straight lines, but are always confined within a domain boundary that separates anti-phase ground-state domains made of type-I vertices (Fig. 4b). Moreover, these monopoles are not free particles, but topological defects that allow antiferromagnetic domains to grow; that is, they are charged objects embedded within a magnetically ordered state.

Statistical ensemble. To characterize the statistical ensemble of monopoles in the square-ice regime, we compared the mean value of the pairing length of the monopole/anti-monopole pairs (Extended Data Fig. 6) and the correlation length ξ deduced from the pinch-point analysis. We find that $\xi = 4.4a$, whereas the mean pairing length is about $3a$. These lengths are comparable and we can consider that the finite width of the pinch points is related to the presence of topological excitations within the Coulomb phase. Consequently, each monopole is expected to diffuse almost independently of its counterpart, allowing us to approximate each pair as two Ising domain walls propagating randomly within their one-dimensional chain—their classical Dirac string. Because domain walls are the source of decorrelation in a one-dimensional short-range Ising chain, we can extract an effective temperature $T_{\text{eff}}/J_{\text{nn}}$ by fitting the mean pairing length to the correlation length of the one-dimensional Ising chain $\xi = |\ln[\tanh(T_{\text{eff}}/J)]|$, which gives $T_{\text{eff}}/J \approx 1.15$. This effective temperature should be taken with care and should not be used to show that our demagnetized samples are thermalized. Nevertheless, it is interesting that these quantitative values are compatible with one another as well as with classical Monte Carlo simulations of the associated spin model³⁹.

Thermally active square ice. If the fabrication of thermally active, shifted square ice is the next step, the design we propose here might not be ideal as a way of accessing the superparamagnetic regime or in experiments that require high-temperature annealing. Using nanometre-thick nanomagnets could enable the superparamagnetic regime to be reached, but the surface of the Ti/Au bases we used might be too rough to allow the growth of a continuous, flat, nanometre-thick permalloy film. In addition, Ti and Au are probably not the best choice, because permalloy does not easily wet on them. For experiments that require thick films to be annealed above the Curie point of the ferromagnet, Ti and Au might again not be the right combination of materials because interdiffusion and dewetting might occur upon annealing at several hundreds of degrees Celsius. Simply because of these materials issues, our system needs to be optimized before being used to fabricate thermally active systems. Magnetic imaging is also challenging. Magnetic force microscopy is not the best technique with which to probe small amounts of material, especially in the superparamagnetic regime, because tip/sample interaction could affect the measurements. Photoemission electron microscopy is also difficult. One reason for this is that X-rays arrive on the sample at a 16° angle with respect to the surface, leading to shadowing effects in systems with important height profiles^{41–43}. Even though photoemission electron microscopy can probe very small amounts of material, the geometry is not ideal for achieving magnetic contrast or accessing, for example, time evolution of spin-flip events in shifted square lattices.

Data availability. The datasets generated and analysed here are available from the corresponding author on reasonable request.

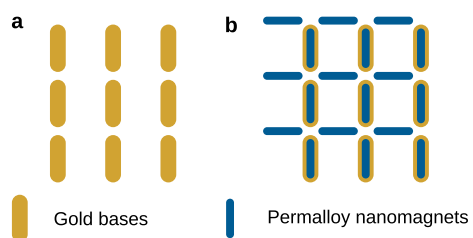
- Rougemaille, N. *et al.* Chiral nature of magnetic monopoles in artificial spin ice. *New J. Phys.* **15**, 035026 (2013).
- Zeissler, K. *et al.* The non-random walk of chiral magnetic charge carriers in artificial spin ice. *Sci. Rep.* **3**, 1252 (2013).
- Donahue, M. J. & Porter, D. G. *OOMMF User's Guide, Version 1.0*. Report No. NISTIR 6376 (National Institute of Standards and Technology, 1999).
- Wang, R. F. *et al.* Demagnetization protocols for frustrated interacting nanomagnet arrays. *J. Appl. Phys.* **101**, 09J104 (2007).
- Morgan, J. P., Bellew, A., Stein, A., Langridge, S. & Marrows, C. H. Linear field demagnetization of artificial magnetic square ice. *Front. Phys.* **1**, 28 (2013).
- Garanin, D. A. & Canals, B. Classical spin liquid: exact solution for the infinite-component antiferromagnetic model on the kagomé lattice. *Phys. Rev. B* **59**, 443–456 (1999).
- Fennell, T. *et al.* Magnetic Coulomb phase in the spin ice $\text{Ho}_2\text{Ti}_2\text{O}_7$. *Science* **326**, 415–417 (2009).
- Isakov, S. V., Moessner, R. & Sondhi, S. L. Why spin ice obeys the ice rules. *Phys. Rev. Lett.* **95**, 217201 (2005).
- Henry, L.-P. *Classical and Quantum Two-dimensional Ice: Coulomb and Ordered Phases*. <https://tel.archives-ouvertes.fr/tel-00932367/document> PhD thesis, Ecole normale supérieure de Lyon (2013).
- Phatak, C., Petford-Long, A. K., Heinonen, O., Tanase, M. & De Graef, M. Nanoscale structure of the magnetic induction at monopole defects in artificial spin-ice lattices. *Phys. Rev. B* **83**, 174431 (2011).
- Kimling, J. *et al.* Photoemission electron microscopy of three-dimensional magnetization configurations in core-shell nanostructures. *Phys. Rev. B* **84**, 174406 (2011).
- Da Col, S. *et al.* Observation of Bloch-point domain walls in cylindrical magnetic nanowires. *Phys. Rev. B* **89**, 180405 (2014).
- Jamet, S. *et al.* Quantitative analysis of shadow x-ray magnetic circular dichroism photoemission electron microscopy. *Phys. Rev. B* **92**, 144428 (2015).



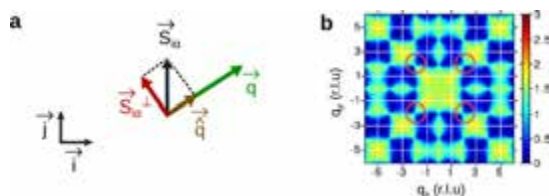
Extended Data Figure 1 | Dumbbell description of the nanomagnets.

Map of J_1/J_2 as a function of l/a and h/a for an isolated vertex.

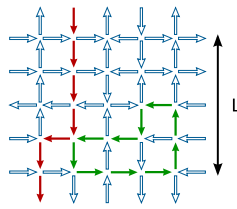
The condition $J_1 = J_2$ is indicated by the dark line. Our results perfectly reproduce those reported in ref. 29. The white dots indicate the values that correspond to the different samples studied here.



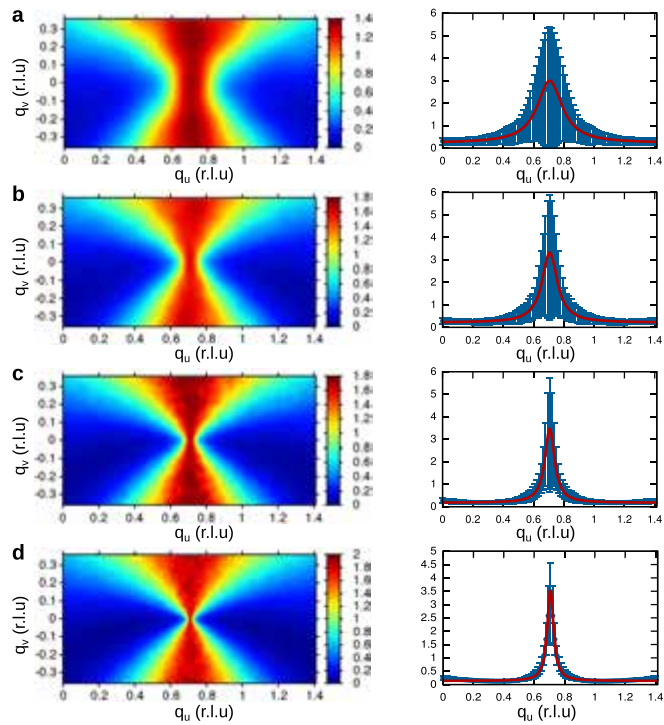
Extended Data Figure 2 | Illustration of the two-step electron-beam lithography process. **a**, Schematic of the gold bases subsequently used to shift the vertical sublattice. **b**, Schematic of the permalloy magnets on the vertical and horizontal sublattices.



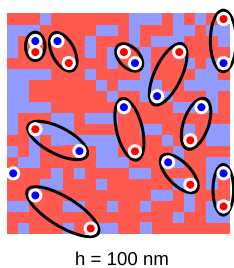
Extended Data Figure 3 | Magnetic structure factor of the square-ice model. **a**, Sketch of the vectors involved in equation (1). **b**, Magnetic structure factor for an ideal square-ice model, computed for 1,000 low-energy states made of $N=840$ spins. Red circles indicate the regions of interest for the intensity profiles in Fig. 3f and Extended Data Fig. 5.



Extended Data Figure 4 | Loop flips in the square lattice. Schematic illustrating the open (red arrows) and closed (green arrows) spin loops used to generate a low-energy configuration that is representative of the massively degenerate ground-state manifold of the square-ice model¹⁹. The lattice contains 840 spins and the number of loops that are flipped between two decorrelated configurations is set to $N = 840$. L corresponds to the linear size of the square lattices.



Extended Data Figure 5 | Analysis of the pinch points. **a–d**, Maps of the pinch points indicated by red circles in Extended Data Fig. 3b (left) and associated intensity profiles along the $q_v = 0$ direction (right), for different lattice sizes L : $L = 10$ (**a**), $L = 20$ (**b**), $L = 40$ (**c**), $L = 80$ (**d**). The colour scale refers to the intensity at a given point of reciprocal space. The coordinates (q_u , q_v) are relative to the intensity profile and do not correspond to the real axes of reciprocal space. The red curves are single-peaked Lorentzian fits; the points represent the mean and the error bars represent the standard deviation calculated from 1,000 random ice-rule configurations.



Extended Data Figure 6 | Magnetic monopoles in artificial square ice. Experimental spin configuration for $h = 100 \text{ nm}$. Type-I and -II vertices appear as blue and red squares, respectively. Monopoles appear as red and blue circles. Their associated pairing is represented by black ellipses.

Extended Data Table 1 | Correlation lengths extracted from the intensity profiles

L	10	20	40	80
ξ	$3.19a \pm 2\%$	$5.2a \pm 4\%$	$8.2a \pm 7\%$	$13a \pm 14\%$

Accessing non-natural reactivity by irradiating nicotinamide-dependent enzymes with light

Megan A. Emmanuel, Norman R. Greenberg, Daniel G. Oblinsky & Todd K. Hyster

Enzymes are ideal for use in asymmetric catalysis by the chemical industry, because their chemical compositions can be tailored to a specific substrate and selectivity pattern while providing efficiencies and selectivities that surpass those of classical synthetic methods¹. However, enzymes are limited to reactions that are found in nature and, as such, facilitate fewer types of transformation than do other forms of catalysis². Thus, a longstanding challenge in the field of biologically mediated catalysis has been to develop enzymes with new catalytic functions³. Here we describe a method for achieving catalytic promiscuity that uses the photoexcited state of nicotinamide co-factors (molecules that assist enzyme-mediated catalysis). Under irradiation with visible light, the nicotinamide-dependent enzyme known as ketoreductase can be transformed from a carbonyl reductase into an initiator of radical species and a chiral source of hydrogen atoms. We demonstrate this new reactivity through a highly enantioselective radical dehalogenation of lactones—a challenging transformation for small-molecule catalysts^{4–7}. Mechanistic experiments support the theory that a radical species acts as an intermediate in this reaction, with NADH and NADPH (the reduced forms of nicotinamide adenine nucleotide and nicotinamide adenine dinucleotide phosphate, respectively) serving as both a photoreductant and the source of hydrogen atoms. To our knowledge, this method represents the first example of photo-induced enzyme promiscuity, and highlights the potential for accessing new reactivity from existing enzymes simply by using the excited states of common biological co-factors. This represents a departure from existing light-driven biocatalytic techniques, which are typically explored in the context of co-factor regeneration^{8,9}.

Living organisms use photoactive co-factors to convert light into chemical energy, driving a litany of biologically essential enzymatic reactions. In these systems, the small-molecule co-factor collects solar energy, while the structure of the associated enzyme dictates how the energy is converted into chemical reactivity. Such specialization of function enables diverse reactivity to be achieved with a small number of photoactive co-factors. Flavin, for example, contributes to the repair of thiamine dimers by one enzyme, the induction of flagellum-mediated locomotion by another, and the setting of circadian rhythms by yet another¹⁰. Yet despite this diversity of function, light-responsive enzymes that act on small organic molecules—generating useful radical species that could drive subsequent reactivity—are rare¹¹. In contrast, in synthetic photoredox catalysis, photon-responsive molecules are commonly used to generate organic radicals from small organic molecules¹². We hypothesized that if enzymes with photoactive co-factors could be adapted to generate radical intermediates, it might be possible to catalyse asymmetric radical-driven reactions.

In designing our system, we sought proteins in which the binding site for organic substrates is adjacent to the binding site for a photoreponsive co-factor, with the expectation that substrate binding might therefore elicit superior selectivity for radical transformations¹³. As regards the co-factor, we were interested in NADH/NADPH, owing to their unique photophysical properties. In its ground state, NADH

(or NADPH) is primarily understood as a hydride (H^-) source and a weak single-electron reductant. But upon photoexcitation it becomes a potent single-electron reductant that can reduce an array of functional groups (Fig. 1a)^{14–17}. For example, the calculated oxidation potential of photoexcited 1-benzyl-1,4-dihydronicotinamide (BNAH^{+*} , an NADH model) for a saturated calomel electrode, a standard reference electrode, is -2.6 V ; BNAH^{+*} also has a low homolytic bond-dissociation free energy (the energy required to break a covalently bonded molecule into two radicals) for the C4–H bond (this energy requirement is 32 kcal mol^{-1} for acetonitrile), thus enabling a thermodynamically favourable transfer of hydrogen atoms (H^\bullet) to most radical species¹⁸.

In our model system, we tested nicotinamide-dependent ketoreductases (KREDs) for their ability to act as an initiator of radical species and a chiral hydrogen-atom source for radical dehalogenation reactions (Fig. 1b)^{19,20}. KREDs are widely used in the production of chemicals, reducing ketones to enantiomerically pure alcohols via hydride transfer (Fig. 1c)²¹. The ubiquity, utility and evolvability of KREDs have made them essential tools in chemical synthesis and, as such, panels of structurally diverse KREDs are commercially available.

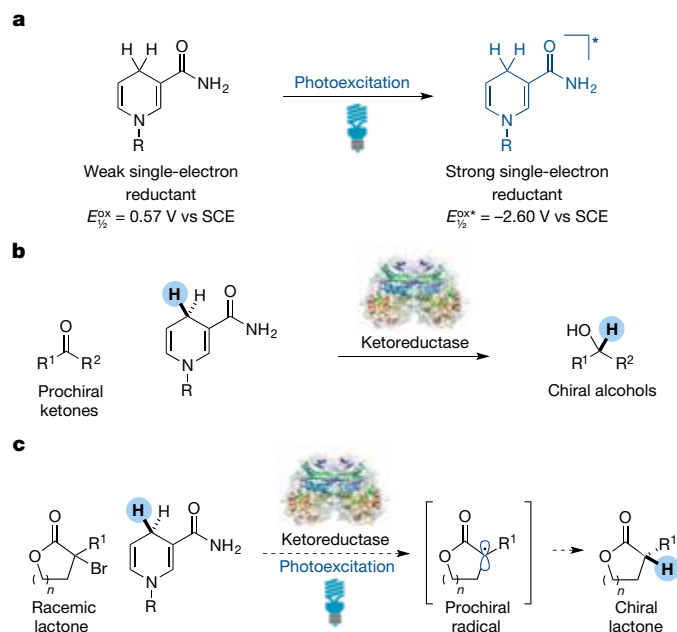
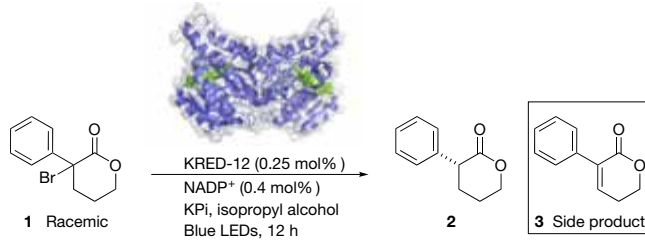


Figure 1 | Photon-induced promiscuity in a biocatalytic process.

a, Ground-state nicotinamide (left) is understood to serve as a hydride (H^-) source and is a weak single-electron reductant. However, when excited (right), it becomes a potent single-electron reductant. $E_{1/2}^{\text{ox}}$, oxidation potential; R, organic functional group; SCE, saturated calomel electrode. **b**, Ketoreductases (KREDs) are highly selective enzymes that catalyse the delivery of hydride from NAD(P)H to prochiral ketones, to access chiral alcohols. **c**, Photoexciting NAD(P)H that is bound to the active site of a KRED should allow for the conversion of racemic halolactones to chiral lactones through an intermediate prochiral radical.

Table 1 | Reaction optimization and control experiments


Entry*	Variations from standard conditions	e.r.	Yield (%)
1	None	98/2	81
2 [†]	Purified enzyme	98/2	85
3	No enzyme	—	<1
4	No light	—	<1
5 [‡]	No NADP ⁺	—	<1
6 [‡]	GDH instead of KRED	—	<1
7 [‡]	LKADH (5 mol%)	—	<1
8 [‡]	LKADH-Y190C (5 mol%)	—	3
9 [‡]	LKADH-E145F-F147L-Y190C (5 mol%)	96/4	72
10 [‡]	SYADH (1 mol%)	37/63	5
11 ^{††}	RasADH (1 mol%)	7.5/92.5	51

Y190C refers to a mutation of the tyrosine amino acid at position 190 in LKADH to cysteine.

E145F, glutamic acid 145 is mutated to phenylalanine. F147L, phenylalanine 147 is mutated to leucine.

*The diagram shows the standard reaction, which was performed at 0.038 mmol scale in 1.25 ml of buffer (made with 1,000 μ l of 125 mM potassium phosphate buffer (KPi) at pH 6.5; 200 μ l isopropyl alcohol; 50 μ l dimethylsulfoxide (DMSO)). Yields were determined by high-performance liquid chromatography (HPLC) relative to an internal standard. Enantioselectivities were determined by chiral HPLC.

[†]Determined using purified protein.

[‡]KRED was replaced with glucose dehydrogenase (GDH).

^{††}Reaction was performed at 0.03 mmol scale in 1.05 ml of buffer (comprising 1,000 μ l of 100 mM TRIS at pH 7.0; 10 mM CaCl₂; 10% glycerol; 200 mM glucose; 50 μ l DMSO) with 5 mg GDH-105.

We selected halogenated lactones as model substrates because they can bind to the active sites of KREDs without being susceptible to carbonyl reduction²².

We tested the viability of the proposed dehalogenation by using a panel of KREDs purchased from Codexis. A halolactone (compound 1) was reacted with a KRED (0.25 mol%) and NADP⁺ (0.4 mol%) in a 20/4/1 mixture of phosphate buffer (125 mM, pH 6.5 with 2 mM MgSO₄)/isopropyl alcohol/dimethylsulfoxide, and then irradiated with blue LEDs (460 nm). Under these conditions, ten KREDs were ineffective, another seven provided the desired lactone (compound 2) at modest yield and variable enantioselectivity, and three variants (KRED-4, KRED-12 and KRED-14) provided lactone 2 with a conversion rate of greater than 95%, a yield of 81%, and an enantiomeric ratio (e.r.) of 98/2 in favour of the (*R*)-enantiomer, with dehydrolactone (compound 3) as the major side product (Table 1, entries 1, 2, and Extended Data Table 1). In suboptimal reactions, the mass balance comprises largely unreacted starting material, with variable amounts of dehydrolactone (3). Notably, compound 3 was completely unreactive under the reaction conditions, eliminating the possibility of a selective alkene reduction (Extended Data Figs 1, 2).

We carried out control experiments in order to elucidate the requirements for this reaction. In the absence of KRED, NADP⁺ or light, we detected unreacted starting material and trace amounts of elimination product (Table 1, entries 3–5). Surprisingly, increasing the concentration of NADP⁺ from 0.4 mol% to 50 mol% produced only a modest decrease in enantioselectivity (Extended Data Fig. 3). Furthermore, when a KRED was replaced with glucose dehydrogenase—in order to turn over NADP⁺ without providing an active site for substrate binding—trace amounts of product were observed (Table 1, entry 6). These results suggest that NADPH that is not bound to an enzyme is far less reactive than bound NADPH. To test this hypothesis, we determined the fluorescence lifetime of NADPH with and without protein: in solution, this lifetime is 0.405 ns; in the protein, the lifetime increases more than 20-fold to 9.00 ns (Extended Data Fig. 5 and Extended Data Table 2)²³. The results of ultraviolet/visible-light experiments suggest that a charge-transfer complex is formed between halolactone (1) and NADPH only in the presence of KRED (Extended Data Fig. 4). It is likely that this complex is responsible for the initial electron transfer. These experiments support the hypothesis

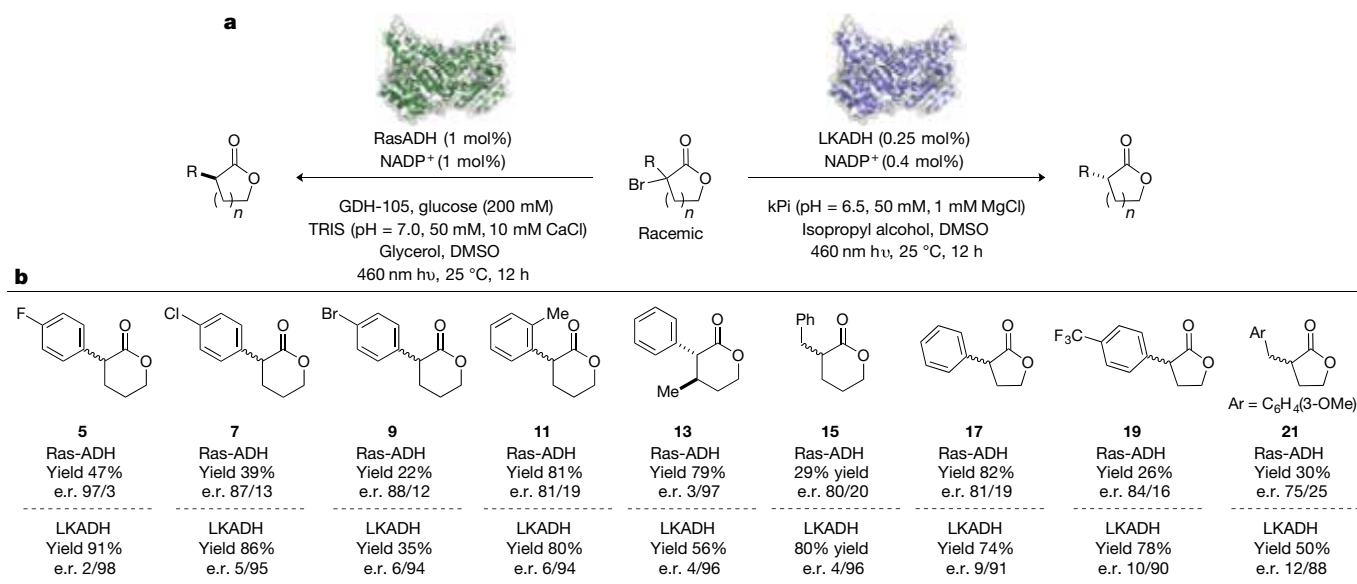


Figure 2 | Substrate scope. An array of different halolactones is amenable to selective dehalogenation activity, providing the (*R*)-enantiomer when using an LKADH variant, or the (*S*)-enantiomer when using RasADH.

a, The basic reactions, starting with a racemic halolactone (centre) and using RasADH (left) or LKADH (right) to catalyse the reactions. The reaction conditions were as follows. For RasADH: halolactone (1 equiv, 40 mM), enzyme (1 mol%), GDH-105 (5 mg), NADP⁺ (1 mol%), glucose (250 mM), TRIS buffer (pH = 7.0, 50 mM, 1 mM CaCl₂, 10% glycerol, 5% dimethylsulfoxide (DMSO), 460 nm, 25 °C, 12 h. For LKADH: halolactone

(1 equiv., 40 mM), enzyme (KRED-12, 0.25 mol%), NADP⁺ (0.25 mol%), KPi buffer (pH = 7.0, 50 mM, 1 mM MgCl₂, 20% isopropyl alcohol, 5% DMSO), 450 nm, 25 °C, 12 h. **b**, The different halolactones that were used as starting points, along with the yields and enantiomeric ratios (e.r.) of the products. For halolactone 13, isotopic labelling suggests that the product rapidly epimerized to the thermodynamically favourable *trans*-product (Supplementary Fig. 9). For halolactones 14, 16, 18 and 20, in the LKADH experiment, KRED-3 was used instead of KRED-12.

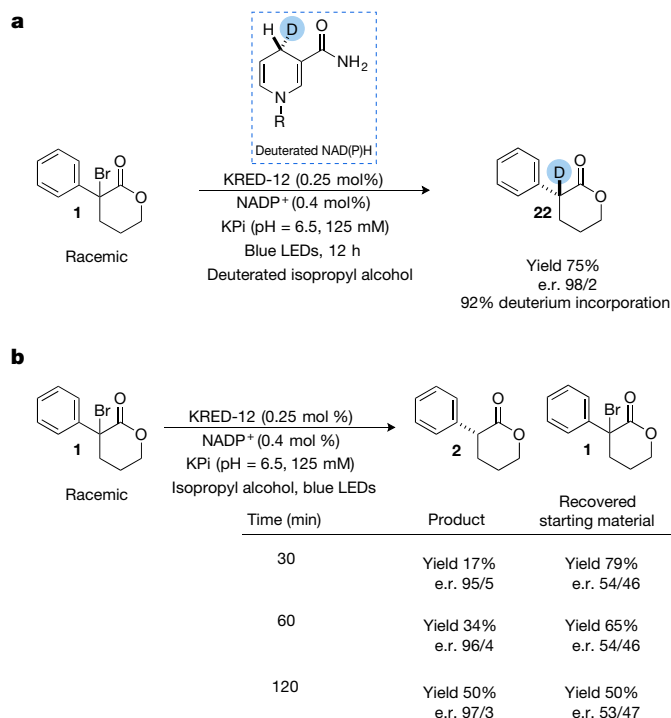
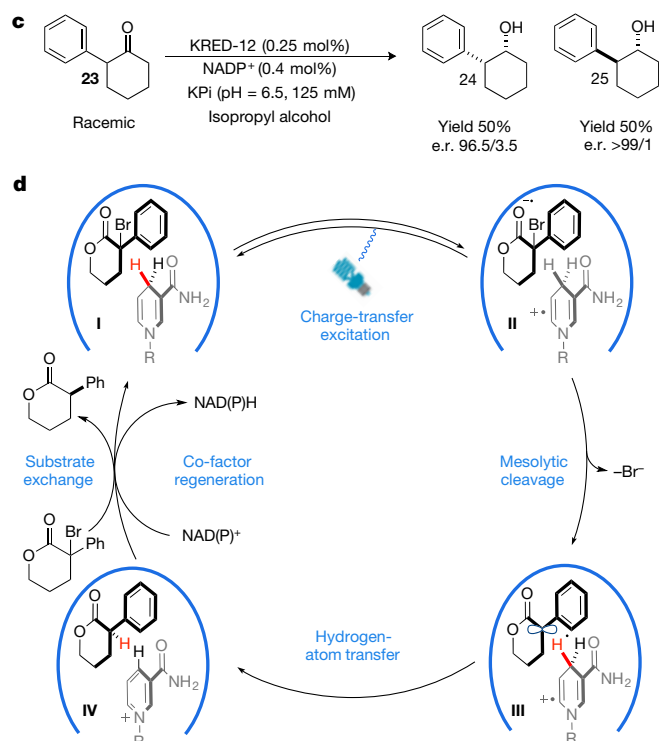


Figure 3 | Mechanistic experiments. **a**, After generating deuterated (D) NAD(P)H *in situ* from deuterated isopropyl alcohol, and under KRED-mediated catalysis, the predominant product is deuterolactone (**22**), showing that NAD(P)H is the hydrogen-atom source in the reaction. **b**, At the end of each reaction, neither enantiomer of starting material is recovered in preference to the other, suggesting that neither enantiomer is preferred by the enzyme. **c**, The most reactive variant of KRED-12 is reactive on both enantiomers of structurally related ketone **23**, supporting the idea that KRED-12 can effect reactions on bulky substrates. **d**, In our proposed mechanism for enantioselective radical dehalogenation, the charge-transfer complex (**I**) comprises either enantiomer of halolactone



(bold) plus NAD(P)H (grey) within the active site of KRED (blue curve). Photoexcitation effects an electron transfer to form complex **II**, in which the halolactone and NAD(P)H are both radicals (denoted by \cdot). Upon mesolytic cleavage of the C–Br bond (that is, dehalogenation), complex **III** is formed, in which the halolactone is a prochiral radical. Then, a conformational change within the enzyme's active site enables hydrogen-atom transfer to the lactone to form an enantioselective product within complex **IV**. Finally, NAD(P) $^+$ can be reduced by either isopropyl alcohol (using native alcohol dehydrogenase activity) or glucose dehydrogenase to complete the catalytic cycle.

that NADPH is stabilized by the protein and increases the quantum yield of the reaction, providing a possible rationale for the negligible racemic background reaction.

Building on the results obtained with the commercially available kit, we explored the potential of structurally characterized KREDs to effect this transformation. Personal correspondence with G. Huisman at Codexis revealed that 16 KRED variants have been derived from the short-chain dehydrogenase of the bacterium *Lactobacillus kefir* (LKADH)²⁴. In the absence of mutations, LKADH can reduce 'small-bulky' substrates, such as acetophenone, but is incapable of reducing 'bulky-bulky' ketones, such as hexanophenone (Supplementary Fig. 1). However, upon mutation of tyrosine 190, LKADH gains the ability to reduce more sterically demanding substrates^{25,26}. G. Huisman revealed that the 10 most active of his 16 KRED variants contain a mutation at position tyrosine 190. Capitalizing on this information, we conducted site-saturation mutagenesis at tyrosine 190 of LKADH, and found that mutation of this amino acid to cysteine did indeed activate the protein for dehalogenation activity, albeit with low product yield (Table 1, entries 7, 8). When we also mutated amino acids that line the enzyme's active site, we discovered a variant (in which glutamic acid 145 was changed to phenylalanine, phenylalanine 147 was changed to leucine, and tyrosine 190 was changed to cysteine) that provided the dehalogenated product at a yield of 72% and an e.r. of 96/4 (Table 1, entry 9).

We hypothesize that these three latter mutations might result in a larger active site. Therefore, wild-type KREDs with naturally large active sites might also be capable of effecting the desired dehalogenation activity. We thus selected short-chain dehydrogenases with large

active sites from the bacteria *Sphingomonas yanoikuyae* (SYADH) and *Ralstonia* species (RasADH)²⁷. While SYADH provided product with a yield of only 5% and an e.r. of 63/37, RasADH was highly effective, providing lactone (**2**) with a yield of 51% and an e.r. of 92.5/7.5, favouring the (*S*)-enantiomer (Table 1, entry 11). We modelled lactone (**2**) into the crystal structure of RasADH (Protein DataBank entry 4BMS) to better understand how substrate binding occurs²⁸. In this model, interactions between the carbonyl oxygen and the side chains of tyrosine 150 and serine 137 are observed, consistent with the known mode of binding for ketones²⁹. Furthermore, the distance between the C4 of NADPH and the α -position of the lactone is reasonable for hydrogen-atom transfer (Extended Data Fig. 6).

With effective enzymes identified, we explored the scope and limitations of this method (Fig. 2). In general, RasADH provided lower yields than KRED-12 or KRED-3, presumably because of the decreased stability of RasADH compared with LKADH. Using bromolactone (**1**) or chlorolactone (**1'**) as the starting halolactone produced similar yields and selectivities (Extended Data Fig. 7). Fluoro- or chloro-substituted lactones provided product in high yields and excellent enantioselectivity (Fig. 2, compounds **5**, **7**). In the case of bromolactone (**9**), decreased conversions were observed owing to poor substrate solubility (Fig. 2, compound **9**). *Ortho*-substituents were well tolerated, providing product in good yield and excellent enantioselectivity (Fig. 2, compound **11**). Interestingly, lactones with decreased redox potentials were viable substrates, providing product in good yields and enantioselectivity (Fig. 2, compounds **15**, **21**). Additional stereocentres were tolerated, providing product with excellent selectivity for the *trans*-isomer (Fig. 2, compound **13**). γ -Lactones were also effective substrates, but required

KRED-3 to achieve good levels of enantioselectivity (Fig. 2, compounds 17, 19 and 21).

We also carried out mechanistic experiments to further elucidate the nuances of this reaction. When the reaction is run with deuterated (D_8) isopropyl alcohol, such that deuterated NAD(P)H is generated *in situ*, deuterolactone (**22**) is formed predominantly (92% deuterium incorporation), with excellent enantioselectivity (e.r. = 98/2), supporting nicotinamide's role as the hydrogen-atom source. (Fig. 3a and Supplementary Fig. 10). Over the course of the reaction, racemic halolactones are converted to enantioenriched product. Although the radical species is prochiral, it is unlikely that it survives for long enough to diffuse into the enzyme's active site. As such, we were curious as to whether a kinetic resolution of the starting material occurs over the course of the reaction. Surprisingly, there appears to be very little preference for one enantiomer of starting material over the other (Fig. 3b). To confirm this hypothesis, we reduced the structurally related ketone (**23**) using KRED-12 (Fig. 3c). The resulting alcohols (**24** and **25**) were isolated with a 50% yield of each diastereomer, with an e.r. of 96/4 and 99/1, respectively, with respect to the alcohol stereocentre. In both cases, the alcohol stereocentre was formed selectively as the (*R*)-isomer, matching the observed selectivity in the dehalogenation reaction. These results suggest that KRED-12 cannot distinguish between enantiomers at the α -position of lactones and cyclic ketones³⁰. Indeed, docking models indicate that RasADH can bind both enantiomers of the starting material (Supplementary Fig. 2).

On the basis of these findings, we propose the mechanism outlined in Fig. 3d. Irradiation of the charge-transfer complex, which comprises halolactone (**1**) and NAD(P)H within the active site of KRED (**I**), effects an electron transfer to form $I^{\bullet-} \cdot NADPH^+ \subset KRED$ (**II**, where \subset symbolizes inclusion in the active site). This, upon mesolytic cleavage of the C–Br bond, forms $I^{\bullet-} \cdot NADPH^+ \subset KRED$ (**III**). We hypothesize that both enantiomers of starting material bind within the active site and, upon dehalogenation and formation of prochiral radical I^{\bullet} , undergo a conformational change within the active site for the enantio-determining hydrogen-atom transfer, to form $2 \cdot NADPH^+ \subset KRED$ (**IV**). Finally, $NADPH^+$ can be reduced by either isopropyl alcohol (using native alcohol dehydrogenase activity) or glucose dehydrogenase to complete the catalytic cycle.

In conclusion, we have found that photoexcitation of nicotinamide-dependent enzymes can cause them to become catalytically promiscuous. We anticipate that this strategy will enable many radical-mediated reactions to be rendered highly selective.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Data Availability The data that support the findings in this study are available from the corresponding author upon reasonable request.

Received 17 May; accepted 14 October 2016.

- Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
- Bornscheuer, U. T. & Kazlauskas, R. J. In *Enzyme Catalysis in Organic Synthesis*, Ch. 41 (eds Drauz, K., Gröger, H. & May, O.) 1695–1723 (Wiley VCH, 2012).
- Prier, C. K. & Arnold, F. H. Chemomimetic biocatalysis: exploiting the synthetic potential of cofactor-dependent enzymes to create new catalysts. *J. Am. Chem. Soc.* **137**, 13992–14006 (2015).
- Blumenstein, M., Schwarzkopf, K. & Metzger, J. O. Enantioselective hydrogen transfer from a chiral tin hydride to a prochiral carbon-centered radical. *Angew. Chem. Int. Edn* **36**, 235–236 (1997).
- Zimmerman, J. & Sibi, M. P. Enantioselective radical reactions. *Top. Curr. Chem.* **263**, 107–162 (2006).
- Meggers, E. Asymmetric catalysis activated by visible light. *Chem. Commun. (Camb.)* **51**, 3290–3301 (2015).
- Frey, P. A. Radical mechanisms of enzymatic catalysis. *Annu. Rev. Biochem.* **70**, 121–148 (2001).
- Maciá-Agulló, J. A., Corma, A. & Garcia, H. Photobiocatalysis: the power of combining photocatalysis and enzymes. *Chemistry* **21**, 10940–10959 (2015).
- Park, J. H. *et al.* Cofactor-free light-driven whole-cell cytochrome P450 catalysis. *Angew. Chem. Int. Edn* **54**, 969–973 (2015).

- Conrad, K. S., Manahan, C. C. & Crane, B. R. Photochemistry of flavoprotein light sensors. *Nat. Chem. Biol.* **10**, 801–809 (2014).
- Gabruk, M. & Mysia-Kurdiel, B. Light-dependent protochlorophyllide oxidoreductase: phylogeny, regulation, and catalytic properties. *Biochemistry* **54**, 5255–5262 (2015).
- Prier, C. K., Rankic, D. R. & MacMillan, D. W. C. Visible light photoredox catalysis with transition metal complexes: applications in organic synthesis. *Chem. Rev.* **113**, 5322–5363 (2013).
- Gu, Y., Ellis-Guardiola, K., Srivastava, P. & Lewis, J. C. Preparation, characterization, and oxygenase activity of a photocatalytic artificial enzyme. *ChemBioChem* **16**, 1880–1883 (2015).
- Fukuzumi, S., Hironaka, K. & Tanaka, T. Photoreduction of alkyl halides by an NADH model compound. an electron transfer chain mechanism. *J. Am. Chem. Soc.* **105**, 4722–4727 (1983).
- Fukuzumi, S., Inada, S. & Suenobu, T. Photoinduced mechanisms of electron-transfer oxidation of NADH analogues and chemiluminescence. Detection of the keto and enol radical cations. *J. Am. Chem. Soc.* **125**, 4808–4816 (2003).
- Jung, J., Kim, J., Park, G., You, Y. & Cho, E. J. Selective debromination and α -hydroxylation of α -bromo ketones using Hantzsch esters as photoreductants. *Adv. Synth. Catal.* **358**, 74–80 (2016).
- Xu, H.-J., Liu, Y.-C., Fu, Y. & Wu, Y.-D. Catalytic hydrogenation of α,β -epoxy ketones to form β -hydroxyketones mediated by an NADH coenzyme model. *Org. Lett.* **8**, 3449–3451 (2006).
- Zhu, X.-Q. *et al.* Determination of the C4-H bond dissociation energies of NADH models and their radical cations in acetonitrile. *Chemistry* **9**, 871–880 (2003).
- Narayanan, J. M. R., Tucker, J. W. & Stephenson, C. R. J. Electron-transfer photoredox catalysis: development of a tin-free reductive dehalogenation reaction. *J. Am. Chem. Soc.* **131**, 8756–8757 (2009).
- Maidan, R. & Willner, I. Photochemical and chemical enzyme catalyzed debromination of meso-1,2-dibromostilbene in multiphase systems. *J. Am. Chem. Soc.* **108**, 1080–1082 (1986).
- Huisman, G. W., Liang, J. & Krebber, A. Practical chiral alcohol manufacture using ketoreductases. *Curr. Opin. Chem. Biol.* **14**, 1–8 (2009).
- Kara, S. *et al.* Access to lactone building blocks via horse liver alcohol dehydrogenase-catalyzed oxidative lactonization. *ACS Catal.* **3**, 2436–2439 (2013).
- Kao, T.-H., Chen, Y., Pai, C.-H., Chang, M.-C. & Wang, A. H.-J. Structure of a NADPH-dependent blue fluorescent protein revealed the unique role of Gly176 on the fluorescence enhancement. *J. Struct. Biol.* **174**, 485–493 (2011).
- Hummel, W. Reduction of acetophenone to *R*(+)-phenylethanol by a new alcohol dehydrogenase from *Lactobacillus kefir*. *Appl. Microbiol. Biotechnol.* **34**, 15–19 (1990).
- Noe, E. L. *et al.* Origins of stereoselectivity in evolved ketoreductases. *Proc. Natl Acad. Sci. USA* **112**, E7065–E7072 (2015).
- Niefind, K., Müller, J., Riebel, B., Hummel, W. & Schomburg, D. The crystal structure of R-specific alcohol dehydrogenase from *Lactobacillus brevis* suggests the structural basis of its metal dependency. *J. Mol. Biol.* **327**, 317–328 (2003).
- Man, H. *et al.* Structures of alcohol dehydrogenases from *Ralstonia* and *Shingobium* spp. reveal the molecular basis for their recognition of 'bulky-bulky' ketones. *Top. Catal.* **57**, 356–365 (2014).
- Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
- Sanli, G., Dudley, J. I. & Blaber, M. Structural biology of the aldo-keto reductase family of enzymes: catalysis and cofactor binding. *Cell Biochem. Biophys.* **38**, 79–101 (2003).
- Cuetos, A. *et al.* Access to enantiopure α -alkyl- β -hydroxy esters through dynamic kinetic resolutions employing purified/overexpressed alcohol dehydrogenases. *Adv. Synth. Catal.* **354**, 1743–1749 (2012).

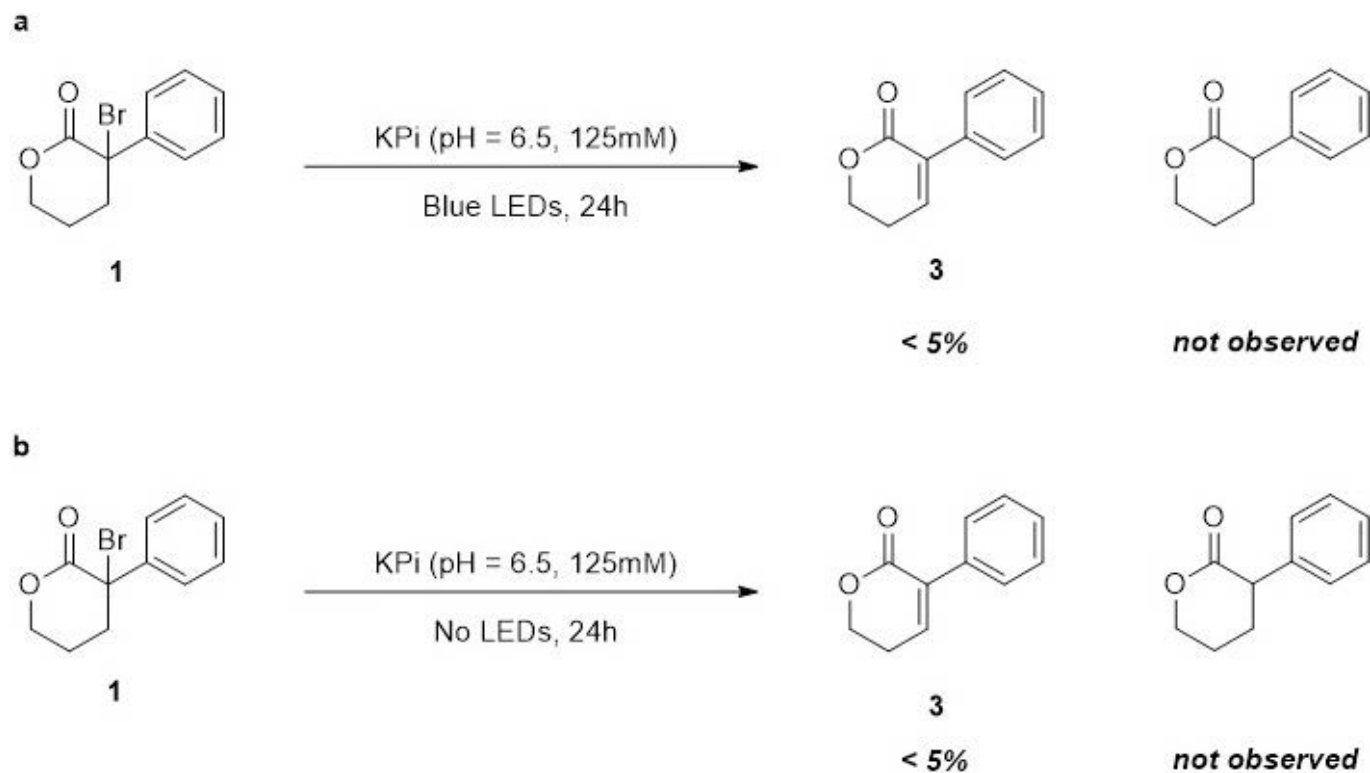
Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support was provided by Princeton University. D.G.O. also acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC). We thank the MacMillan group for use of their chiral high-performance liquid-chromatography and cyclic-voltammetry equipment; G. Scholes for providing the time-resolved fluorescence instrument; H. Yayla of the Knowles group for assistance with cyclic-voltammetry experiments; B. Shields of the Doyle group and the Scholes Group for collection of the LED emission spectrum; and G. Huisman of Codexis for conversations regarding the nature of the mutants in the Codexis KRED kit.

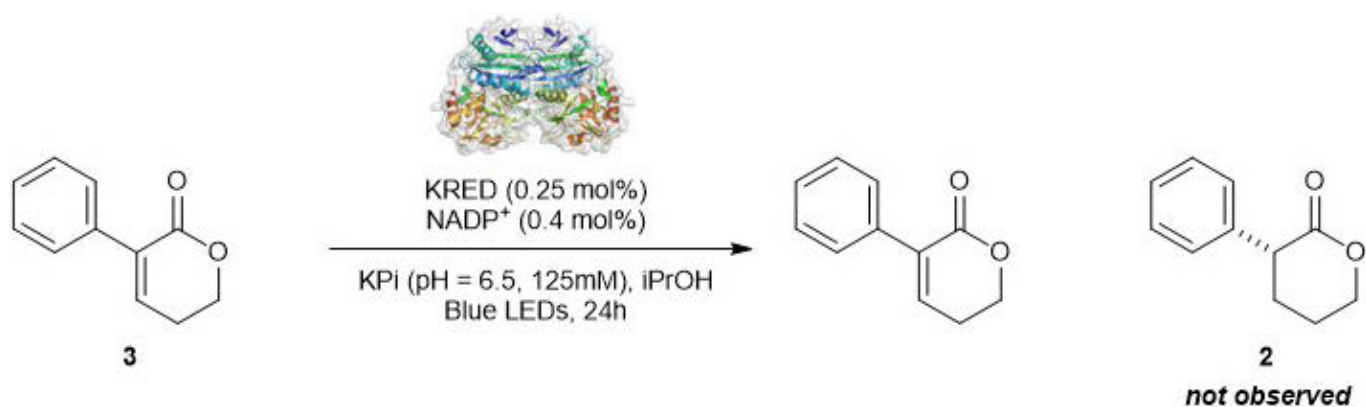
Author Contributions M.A.E. and T.K.H. designed the experiments, performed and analysed experiments, and prepared the manuscript. N.R.G. performed and analysed experiments. D.G.O. collected and analysed time-resolved fluorescence data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.K.H. (thyster@princeton.edu).

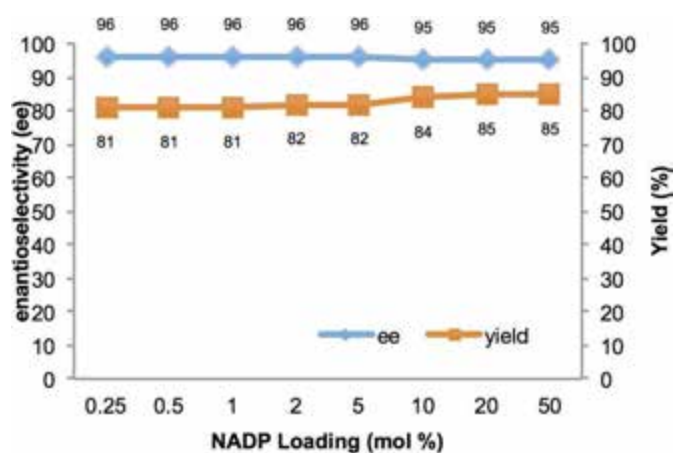
Reviewer Information Nature thanks E. Meggers and the other anonymous reviewer(s) for their contribution to the peer review of this work.



Extended Data Figure 1 | Control experiments for degree of elimination product. **a, b**, Lactone (**1**) does not undergo spontaneous dehydrogenation to produce dehydrolactone (**3**) in solution, either under irradiation (**a**) or without irradiation (**b**). This further suggests that the reaction is not proceeding via KRED-catalysed alkene reduction.

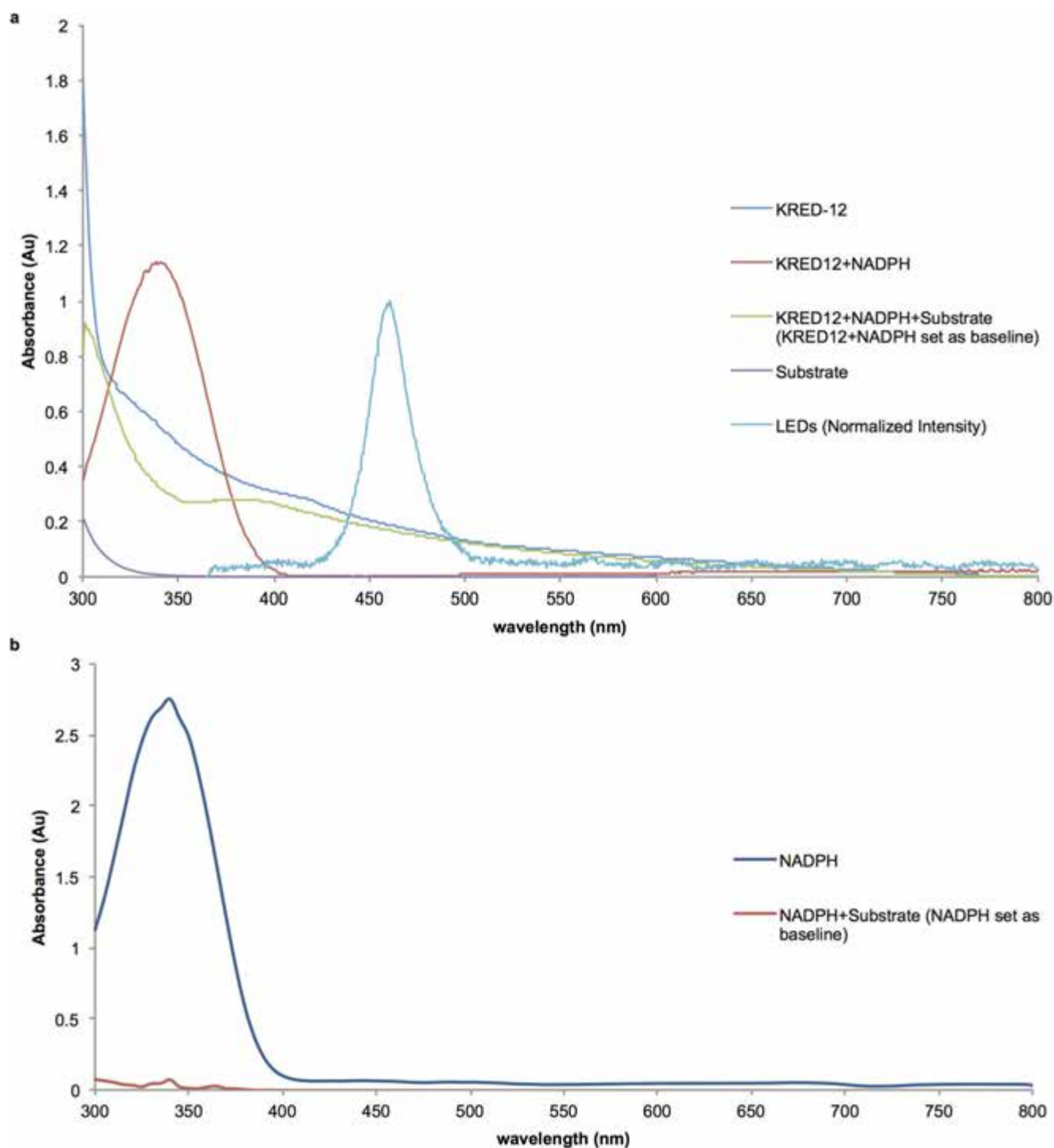


Extended Data Figure 2 | Control for promiscuous alkene-reductase activity. Dehydrolactone (**3**, left) is unreactive under our reaction conditions; thus, product **2** is not formed by KRED-catalysed alkene reduction.



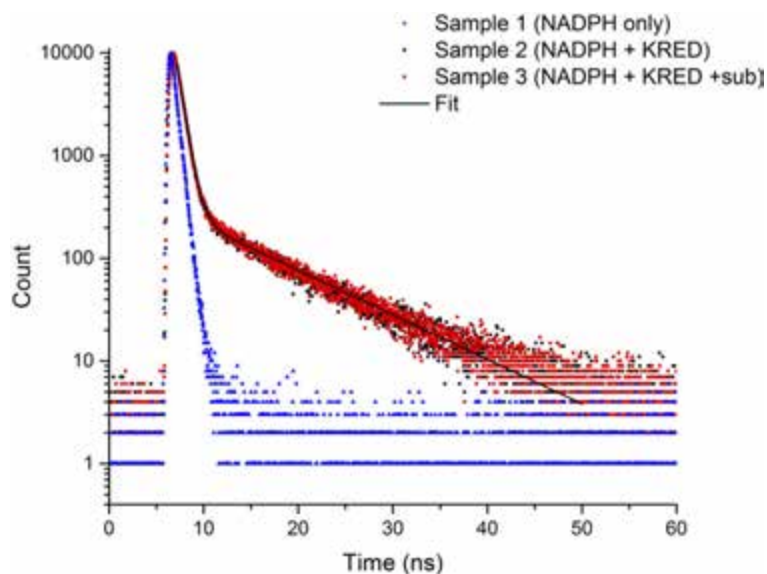
Extended Data Figure 3 | Control for the effects of excess nicotinamide.

Product yield and enantiomeric excess remain relatively unchanged over a wide range of NADP loadings. This suggests that, once KRED is fully loaded with NADP, further increases in NADP concentration have little effect; that is, NADP needs to be bound to KRED in order to exert its effect in this reaction.



Extended Data Figure 4 | Absorbance spectra. a, In the presence of KRED, there is a redshift in the absorbance spectrum of NADPH when substrate is added to the system. This shift is potentially indicative of a charge-transfer complex being formed between the substrate and NADPH. These spectra are overlaid with the emission spectrum of the

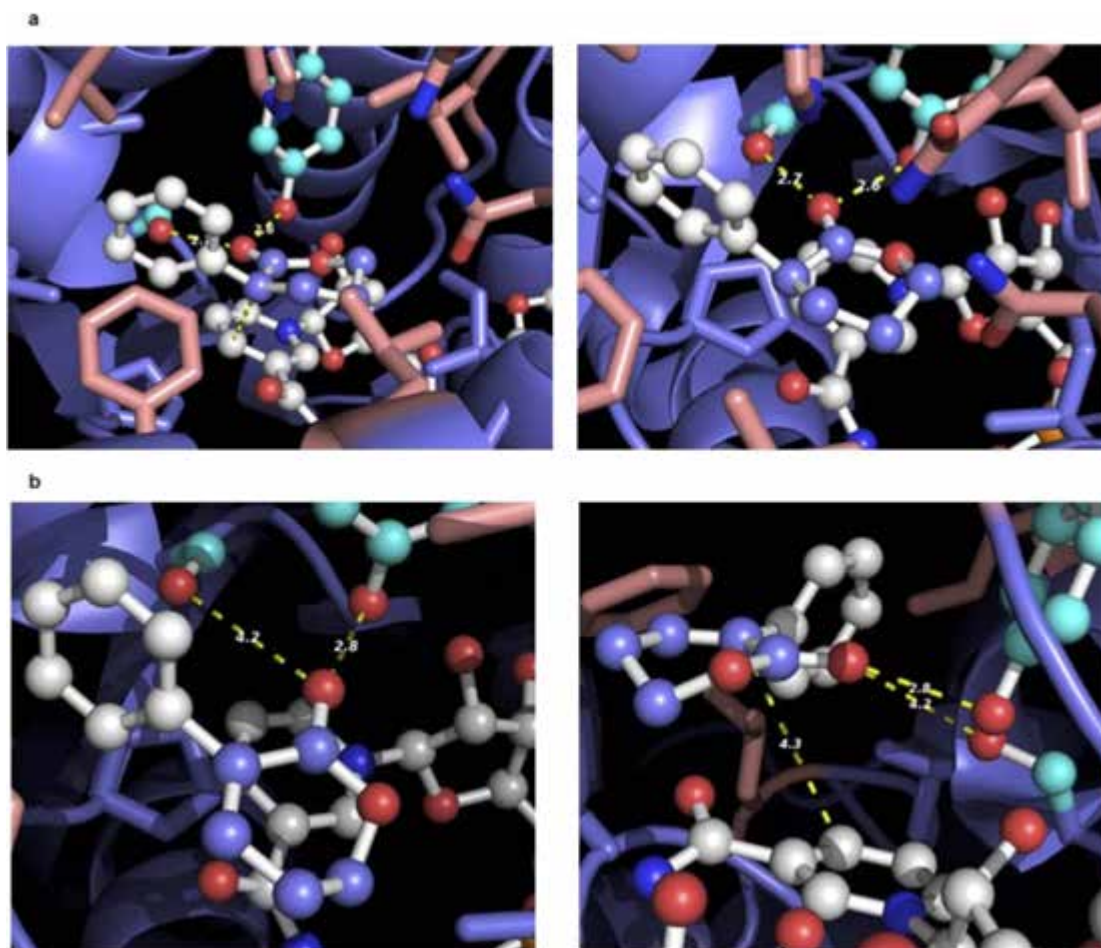
blue LED source used in all experiments. **b,** In the absence of KRED, no such absorption shift is seen when substrate is added. Together, these data suggest that light (LED) irradiates a charge-transfer complex comprising substrate, NADPH and enzyme to initiate the catalytic cycle.



Extended Data Figure 5 | Time-correlated single-photon counting.

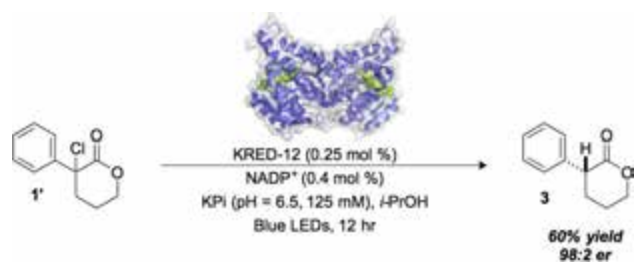
The lifetimes of fluorescence-excited states of NADPH, with and without enzyme and substrate, were determined using time-correlated single-photon counting (TCSPC) on a HORIBA Scientific DeltaFlex TCSPC system. Each sample was concentrated to an optical density of

approximately 0.1 absorbance units and excited using a 305-nm laser; fluorescence emission decay was then probed at 460 nm. Data analysis was done using HORIBA Scientific DAS6 decay analysis software, whereby each data set was fit to an exponential curve to obtain the lifetimes.



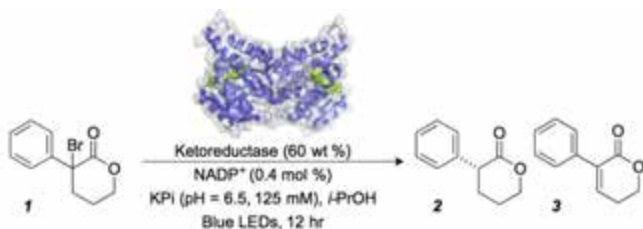
Extended Data Figure 6 | Docking models. Automated docking models were obtained using Autodock Vina²⁸. Panels **a** and **b** indicate two different docking poses as predicted by Autodock. Chain C of RasADH (PDB = 4BMS) was used as a receptor and prepared using Autodock tools. Coordinates for all ligands were prepared using Gaussian and Autodock tools. The active site of RasADH was contained in a $10 \times 10 \times 10$ grid with 1 Å spacing, centred around position $45 \times 9.431 \times 37.226$, which approximately corresponds to the C4–H bond of NADPH.

The exhaustiveness parameter was set to 10 and the rest of the docking parameters were set to default. Docking models for substrate **2** were accessed for their ability to rationalize the observed stereochemistry and provide a reasonable distance and geometry for hydrogen-atom transfer. The blue ribbon shows chain C of RasADH. Substrate **2** is shown in ball-and-stick format, with oxygens in red, lactone carbons in blue, and aromatic carbons in white. The left and right images are two different views of the same model.



Extended Data Figure 7 | Chlorolactone **1' in dehalogenation.** The halolactone **1'** shows similar reactivity to its bromolactone counterpart (**1**) under the same reaction conditions.

Extended Data Table 1 | Screen of commercially available KREDs



Abbreviation	Ketoreductase Variant	Parent ADH	Mutation to Y190	er 2	Yield (%) 2	Yield (%) 3
KRED-1	KRED-P1A04	<i>L. kefir</i>	No	-	8	20
KRED-2	KRED-P1A12	<i>L. kefir</i>	No	-	9	20
KRED-3	KRED-P1B02	<i>L. kefir</i>	Yes	97:3	34	14
KRED-4	KRED-P1B05	<i>L. kefir</i>	Yes	98:2	81	5
KRED-5	KRED-P1B10	<i>L. kefir</i>	Yes	85:15	34	15
KRED-6	KRED-P1B12	<i>L. kefir</i>	Yes	90:10	20	13
KRED-7	KRED-P1C01	<i>L. kefir</i>	Yes	93:7	67	17
KRED-8	KRED-P1H08	<i>L. kefir</i>	No	-	6	15
KRED-9	KRED-P2B02	<i>L. kefir</i>	Yes	97:3	50	8
KRED-10	KRED-P2C02	<i>L. kefir</i>	Yes	93:7	35	8
KRED-11	KRED-P2C11	<i>L. kefir</i>	No	-	7	41
KRED-12	KRED-P2D03	<i>L. kefir</i>	Yes	98:2	81	4
KRED-13	KRED-P2D11	<i>L. kefir</i>	Yes	86:14	25	16
KRED-14	KRED-P2D12	<i>L. kefir</i>	Yes	98:2	81	5
KRED-15	KRED-P2G03	<i>L. kefir</i>	No	-	11	27
KRED-16	KRED-P2H07	<i>L. kefir</i>	No	-	7	25
KRED-17	KRED-P3B03	<i>Thermobacter b.</i>	-	-	7	25
KRED-18	KRED-P3G09	<i>Thermobacter b.</i>	-	-	7	25
KRED-19	KRED-P3H12	<i>Thermobacter b.</i>	-	-	6	24
KRED-20	KRED-134	wild type	-	-	12	23

Reaction results obtained using 20 genetically different KREDs (purchased from Codexis).

Extended Data Table 2 | Time-correlated single-photon counting

Sample		Curve	T1 (s)	T2 (s)
1	NADPH only	Monoexponential	$4.05 \times 10^{-10} \pm 0.026$	-
2	NADPH with KRED	Biexponential	$3.85 \times 10^{-10} \pm 0.125$	$7.92 \times 10^{-9} \pm 1.7$
3	NADPH with KRED and substrate	Biexponential	$4.00 \times 10^{-10} \pm 0.05$	$9.00 \times 10^{-9} \pm 0.12$
4	KRED only	N/A	N/A	N/A
5	Substrate only	N/A	N/A	N/A

Time-correlated single-photon fluorescence decays for NADPH, NADPH plus KRED, and NADPH plus KRED plus substrate. T1 and T2 are the lifetimes of fitted exponential fluorescence decay kinetics. The fluorescence lifetimes obtained for only NADPH are consistent with literature values²³. Samples 2 and 3 produce biexponential curves, indicating the presence of two distinct fluorescent species. Considering the similarities among all T1 values, we determine the fluorescent species to be free NADPH and KRED-bound NADPH. No fluorescence was observed in enzyme-only or substrate-only controls.

High-resolution mapping of global surface water and its long-term changes

Jean-François Pekel¹, Andrew Cottam¹, Noel Gorelick² & Alan S. Belward¹

The location and persistence of surface water (inland and coastal) is both affected by climate and human activity¹ and affects climate^{2,3}, biological diversity⁴ and human wellbeing^{5,6}. Global data sets documenting surface water location and seasonality have been produced from inventories and national descriptions⁷, statistical extrapolation of regional data⁸ and satellite imagery^{9–12}, but measuring long-term changes at high resolution remains a challenge. Here, using three million Landsat satellite images¹³, we quantify changes in global surface water over the past 32 years at 30-metre resolution. We record the months and years when water was present, where occurrence changed and what form changes took in terms of seasonality and persistence. Between 1984 and 2015 permanent surface water has disappeared from an area of almost 90,000 square kilometres, roughly equivalent to that of Lake Superior, though new permanent bodies of surface water covering 184,000 square kilometres have formed elsewhere. All continental regions show a net increase in permanent water, except Oceania, which has a fractional (one per cent) net loss. Much of the increase is

from reservoir filling, although climate change¹⁴ is also implicated. Loss is more geographically concentrated than gain. Over 70 per cent of global net permanent water loss occurred in the Middle East and Central Asia, linked to drought and human actions including river diversion or damming and unregulated withdrawal^{15,16}. Losses in Australia¹⁷ and the USA¹⁸ linked to long-term droughts are also evident. This globally consistent, validated data set shows that impacts of climate change and climate oscillations on surface water occurrence can be measured and that evidence can be gathered to show how surface water is altered by human activities. We anticipate that this freely available data will improve the modelling of surface forcing, provide evidence of state and change in wetland ecotones (the transition areas between biomes), and inform water-management decision-making.

Between any two points in time, part of the Earth's surface is constantly underwater and part is never underwater, with the remainder fluctuating between these extremes. Coastlines and lake and river boundaries advance and retreat, rivers meander, new permanent lakes form and

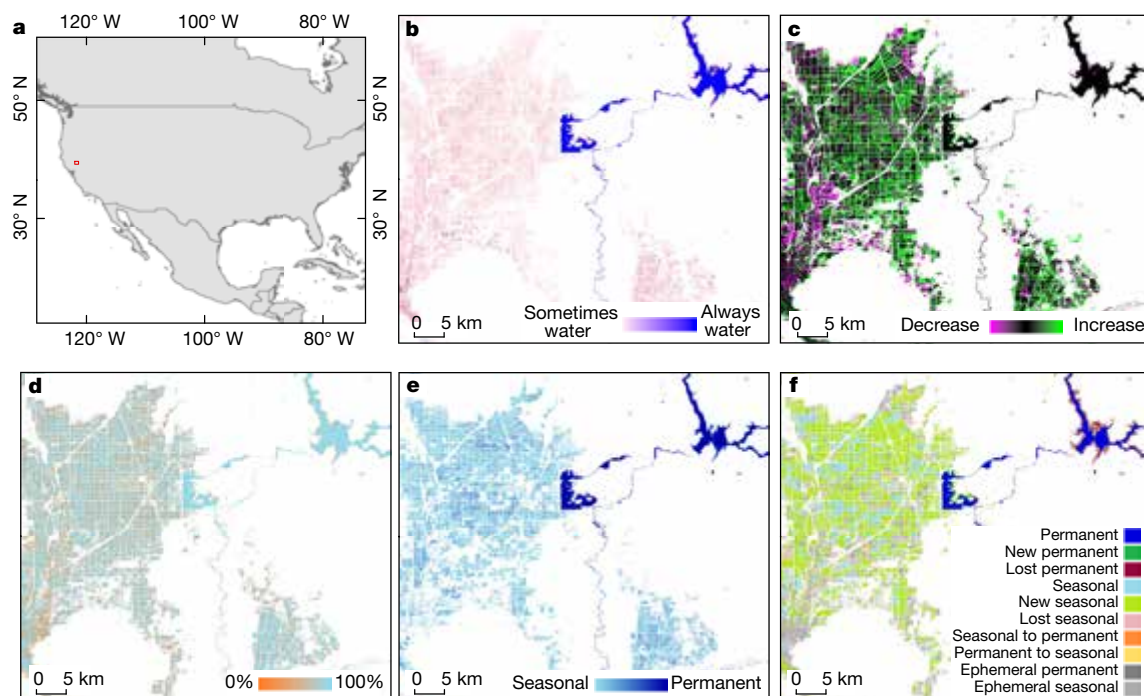


Figure 1 | Different facets of surface water dynamics. **a**, Map of the USA showing Sacramento Valley location (red square). **b**, Surface water occurrence 1984–2015. **c**, Surface water occurrence change intensity 1984–2015. **d**, Surface water recurrence 1984–2015. **e**, Surface water seasonality 2014–2015. **f**, Transitions in surface water class 1984–2015. The Sacramento Valley is one of the major rice-growing

regions in the USA, extracted from the global data set. Seasonal water areas in the left and lower right of each panel correspond to flood irrigation, mainly rice paddies. The more permanent water features (centre and top right of each panel) are reservoirs. See Supplementary Information for a description of the water classes.

¹European Commission, Joint Research Centre, Directorate for Sustainable Resources, 20127 Ispra, Lombardy, Italy. ²Google Switzerland GmbH, Brandschenkestrasse 110, 8002 Zürich, Switzerland.

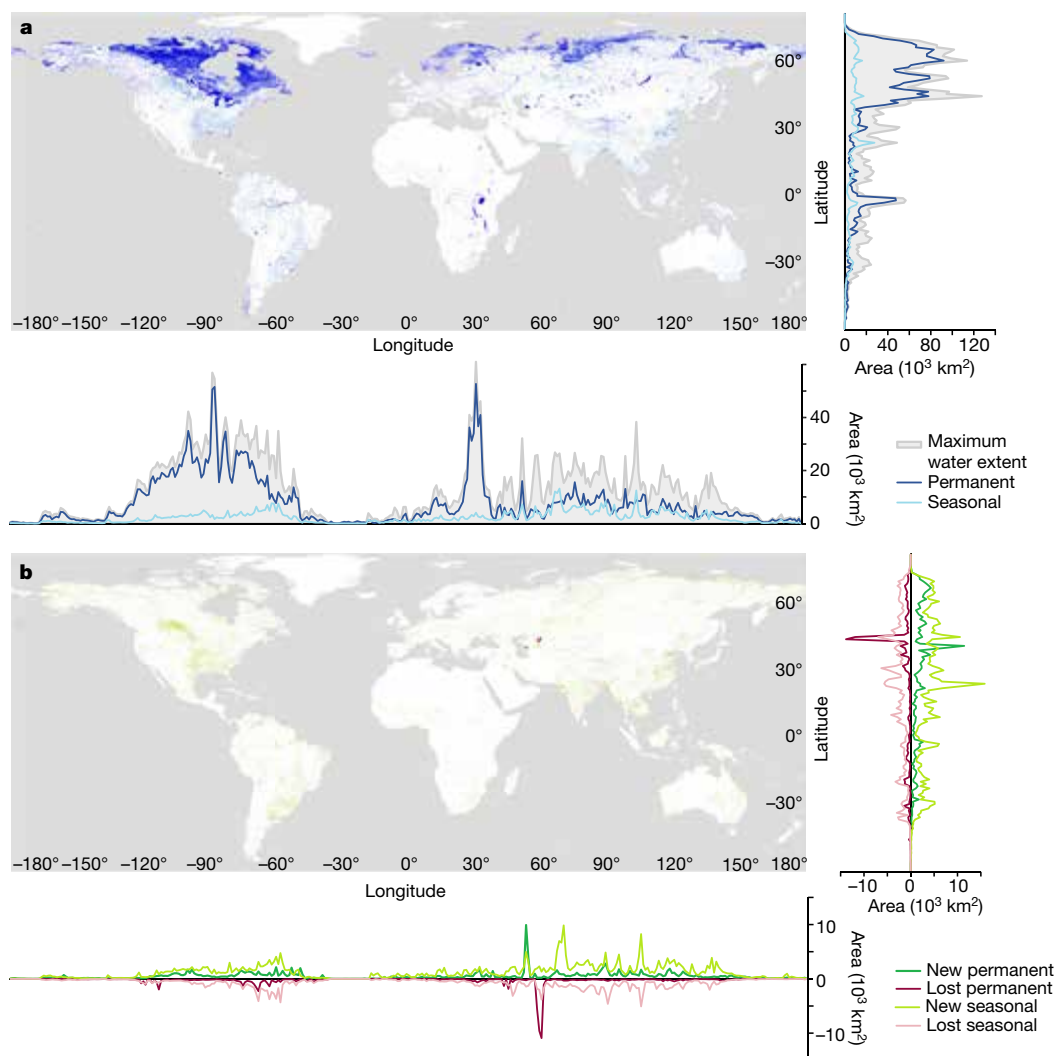


Figure 2 | Global surface water distribution and changes. Global maps, with 1° latitude/longitude summaries of surface water area shown on the right and underneath. **a**, Maximum water extent, permanent and seasonal surface water occurrence October 2014 to October 2015. **b**, Gains and loss

in permanent and seasonal surface water area between 1984 and 2015. All measurements made from inland and coastal waters are defined only by the GADM reference layer (see Methods).

others empty, while seasonal inundations and flood-irrigation cycles periodically create temporary bodies of water. When and where you find water on the planet's surface is hugely important. The presence (or absence) of water influences the climate system, as accounted for in general circulation models², as well as CO₂ evasion and methane emissions³. Access to water influences movement, viable range and migrations for multitudes of species⁴; it is indispensable for sustainable development⁵ and can threaten the security of people, institutions and economies⁶.

Global surface water dynamics have been recorded from coarse-spatial-resolution satellite observations¹², higher-resolution seasonality maps have been produced using Landsat satellite imagery at 5- to 10-year intervals¹¹, and all Landsat images over multiple decades have been used to map seasonality and changes at continental¹⁹ and sub-continental²⁰ scales. The data set presented here (freely available from <https://global-surface-water.appspot.com/>) extends previous work by using the entire multi-temporal orthorectified Landsat 5, 7 and 8 archive spanning the past 32 years to map the spatial and temporal variability of global surface water and its long-term changes.

Each pixel in 1,823 terabytes of Landsat data (Extended Data Fig. 1a–c) was classified as open water, as land or as a non-valid observation using an expert system (see Methods, Extended Data Figs 2 and 3, and Supplementary Table 2). Open water is any stretch of water larger than 30 m by 30 m open to the sky, including fresh and

saltwater. Classification performance, measured using over 40,000 reference points (Methods and Extended Data Figs 4 and 5) confirmed that the classifier produces less than 1% of false water detections, and misses less than 5% of water (Extended Data Table 1).

The long-term water history was used to produce thematic products that document different facets of surface water dynamics. Figure 1 shows extracts from the global products for part of the USA's Sacramento Valley (see Fig. 1a). Figure 1b shows occurrence (variations in persistence and location) between March 1984 and October 2015. The intensity with which occurrence increased or decreased over the 32 years documents gain, loss and constancy in persistence (Fig. 1c). The frequency with which water reappears from year to year across the time-series is mapped as recurrence (Fig. 1d), and water surfaces present throughout an entire year's observations are separately mapped from those that are seasonal (Fig. 1e). Transitions between permanent water, seasonal water and land classes can be determined between any two years of observation; transitions between the first and last year of observation are shown in Fig. 1f. Temporal profiles document water history per pixel, per month and year, and change measurements at the global, continental and country scales are produced by combining these complementary information layers (Supplementary Table 1).

Three per cent (4.46 million km²) of the Earth's landmass was under water at some time between March 1984 and October 2015. Figure 2a

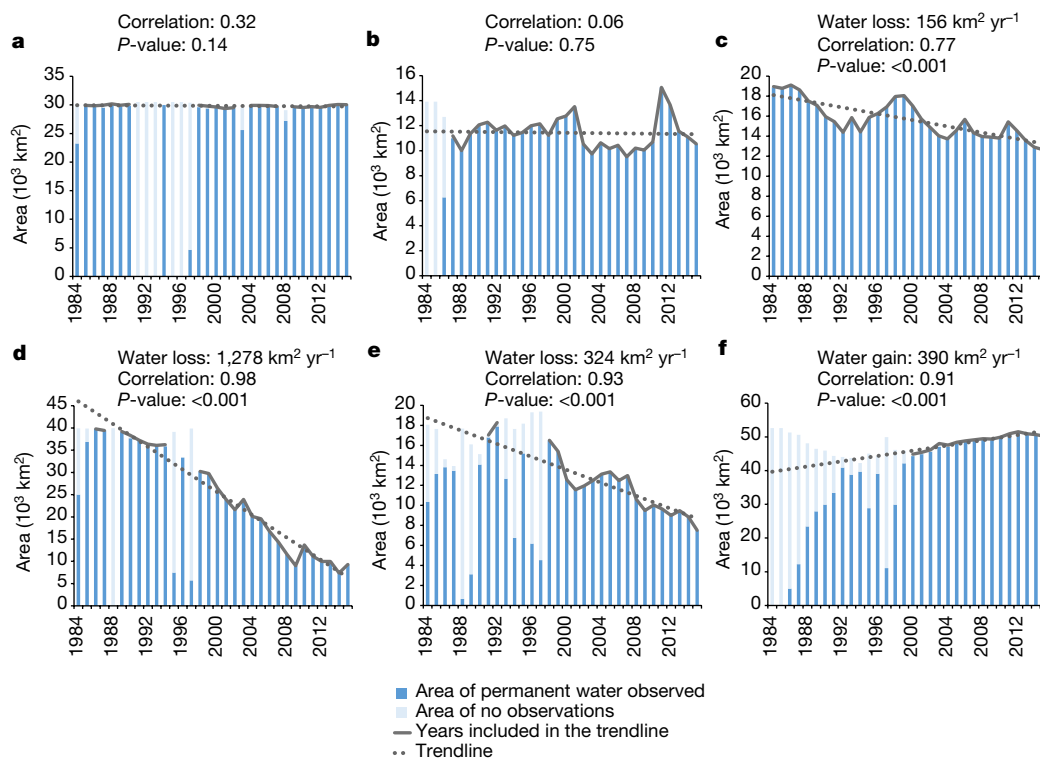


Figure 3 | Trends in annual permanent water surface area. **a**, Finland. **b**, New South Wales, Australia. **c**, Western states of the USA (Arizona, California, Idaho, Nevada, Oregon, Utah). **d**, Aral Sea (Kazakhstan, Uzbekistan). **e**, Iraq, Iran and Afghanistan. **f**, Tibetan plateau. Uncertainty

is estimated from the unobserved component of the maximum permanent water extent. The true surface water area is within this range. Trend lines are provided from years where the unobserved component is less than 5%.

shows that over half (52%) of this is found above 44° N, a pattern that corresponds overall with previous work^{9–11}. In 2015 permanent bodies of water covered 2.78 million km², and 86% of this (2.4 million km²) was geographically and temporally invariant, being consistently present across the entire observation record; the world's ancient lakes such as Baikal and Tanganyika, North America's Great Lakes and the Nordic region's 'land of a thousand lakes' are part of these truly permanent waters. But Fig. 2b also reveals striking patterns of surface water occurrence change.

Over the past three decades more than 162,000 km² of water bodies previously thought of as permanent have proved not to be so; almost 90,000 km² have vanished altogether and over 72,000 km² have transitioned from a permanent state to a seasonal state. During the same period almost 213,000 km² of new permanent water bodies came into existence; 29,000 km² of these used to be seasonally flooded but are now underwater all year round, while 184,000 km² of permanent water formed in areas previously devoid of surface water. In the contemporary timeframe (October 2014 to October 2015) seasonal water covered 0.81 million km².

Surface water is only a part of the water resource, but it is the most accessible to human populations²¹, and provides wide-ranging ecosystem services. Almost 52% of the planet's truly permanent surface water occurs in North America, home to less than 5% of the population in 2015²², and the continent also holds 18% of the contemporary seasonal water. Between 1984 and 2015 North America's permanent water area increased by 17,000 km². In contrast, Asia, with 60% of the human population, accounts for only 9% of the truly permanent and 35% of the contemporary seasonal water. Asia has gained 71,000 km² of permanent water, which is a 23% increase for the continent. Africa and Latin America have almost the same share of the world's permanent water at around 9%, though their populations are very different, with Africa (16% of the total) supporting nearly twice as many people as Latin America (8.6%). Europe, including Russia, with 10% of the global population, has 22% of the permanent water and 18% of the

contemporary seasonal water. Oceania is the only continental region with a net loss of permanent water, albeit a tiny area at 229 km².

The continental summaries obscure strong regional variation. Australia's millennium drought (between 2001 and 2009) substantially affected hydrology¹⁶, and measurable impacts of this can be seen in the permanent surface water area trends for the drought years (Fig. 3b). Regional variation is also apparent in the USA. Although the country's permanent surface water area overall has increased by 0.5% since 1984, a combination of drought and sustained demands for water¹⁷ have seen six western states lose 33%, more than 6,000 km² (Fig. 3c). These losses are modifying social behaviour, driving local water-management policy changes and causing shifts in agricultural production¹⁷.

Human behaviour may be changing in response to water distribution, but human action is itself changing surface water patterns. Over 70% of global net permanent water loss is concentrated in five countries. The marked negative anomaly in permanent water cover change centred at 45° N, 60° E (Fig. 2b) corresponds to this hotspot of change. Kazakhstan and Uzbekistan have lost much of the eastern lobe of the southern Aral Sea (Figs 3d and 4). The rate of loss was greatest between 1994 and 2009, though lately this has slowed and even partially reversed (Fig. 3d). Diversion of, and withdrawal from, the Amu and Syr rivers that once fed the lake are the main causes of loss, but changes in water management offer hope of stabilization and partial restoration¹⁵. Iran, Afghanistan and Iraq have also undergone major losses, having respectively 56%, 54% and 34% less permanent surface water in 2015 than in the first year of observation (Fig. 3e). These losses, which raise serious questions concerning water security and transboundary water management²³, are caused by factors including unregulated withdrawal, dams that change the flow rate and direction of rivers, and droughts¹⁶.

24 countries spread across all continents have each gained at least 1,000 km² of new permanent bodies of water (Supplementary Table 1). Much comes from reservoir construction (Extended Data Fig. 6),

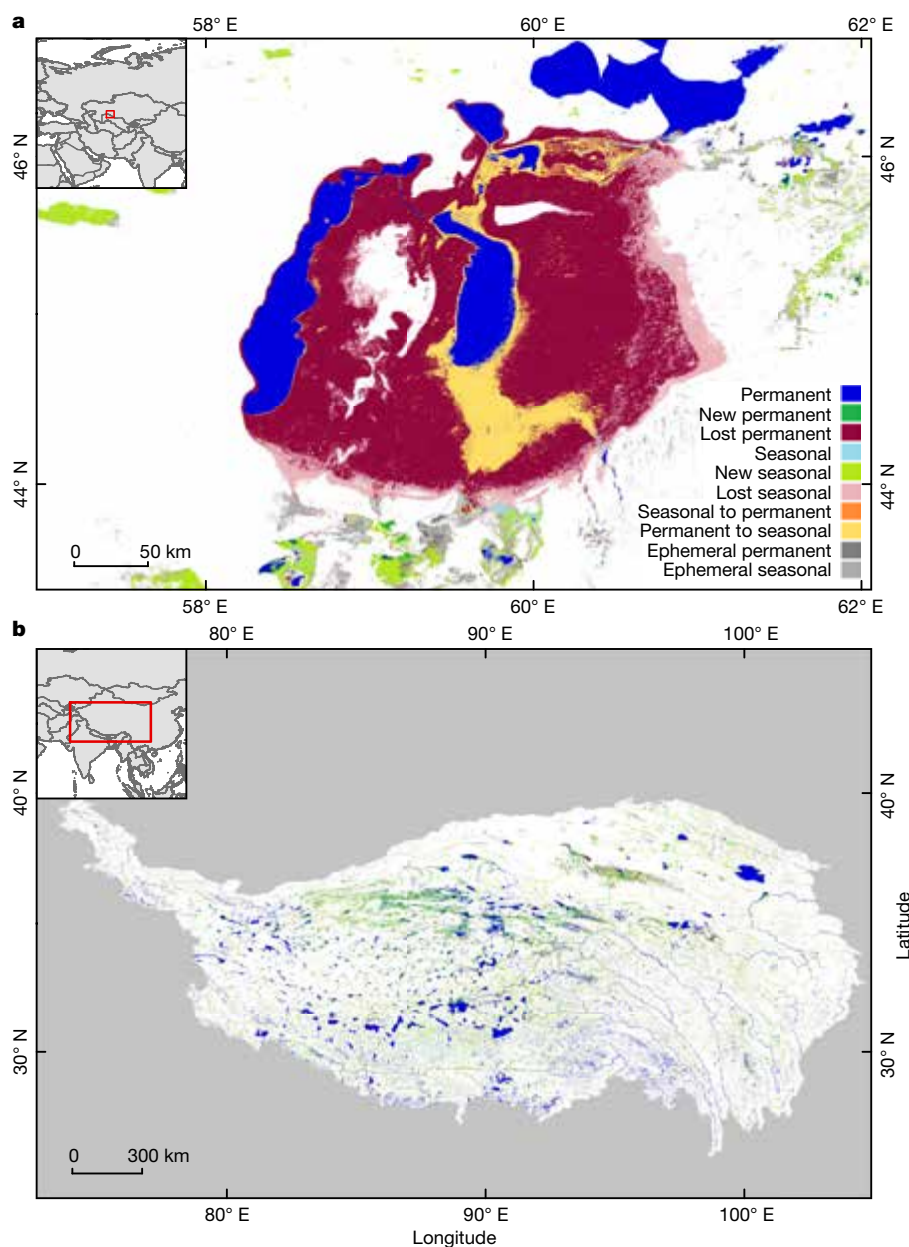


Figure 4 | Surface water changes for the Aral Sea and Tibetan plateau. Transitions between the first year in which representative observations were acquired and the last year of observation. **a**, The Aral Sea. **b**, Tibetan

plateau. The regional context for each location is shown in the insets. The 32-year trends in water surface area for these mapped regions are shown in Fig. 3d and f. The key to the seasonal classes applies to both panels.

with most of the countries featuring prominently in the International Commission on Large Dams register of dam builders²⁴. Turkmenistan is an exception; there over 90% of the additional water surface—14,411 km²—comes from the recovery of the Garabogazköl Aylagy lagoon following the breach in 1992 of a dam between the lagoon and the Caspian Sea²⁵. Permanent surface water has also greatly increased without dam building on the Tibetan plateau (Fig. 4). Virtually all of the region's endorheic lakes are expanding and new lakes are forming, leading to a 20% increase, an extra 8,300 km² (Fig. 3f). Lake expansion on the plateau has been linked to increased run-off from accelerated snow-and-glacier melt caused by higher temperatures and annual precipitation¹⁴. Climate change adaptation challenges include grazing land reduction caused by inundation, grassland degradation linked to salination following the expansion of brackish waters, and threats to transport infrastructure²⁶. The climate-related gains in surface water in this region contrast markedly with the drought-related losses in Australia, the western states of the USA and in central Asia/the Middle East described above.

Seasonal water surfaces can show strong variability, moving between wet and dry years, even shifting geographically. Capturing such variability, especially for short-duration events, is challenging because cloud-free satellite observation must be concurrent with the water occurrence. The accuracy of seasonal water mapping is correspondingly lower (Extended Data Table 1). Changes in location, duration and timing depend on prevailing weather conditions (including changes driven by major perturbations such as the El Niño–Southern Oscillation), though erosion, sediment transport and deposition (particularly around coastlines and along river courses) and land-use choices also have an impact. Changes in any of these conditions can even drive transitions between permanent and seasonal classes (as defined in Methods). For example, many seasonally flooded paddy fields around the Sundarbans mangrove forest in Bangladesh and India have transitioned into permanently inundated fishponds (Extended Data Fig. 7). This may be through choice and market forces but may also be from necessity, as paddy field water and soils become increasingly saline with rising sea

levels and subsiding delta lands²⁷. Seasonally flooded lands can also be drowned in the backwaters of new dams, while downstream fluvial systems become increasingly fragmented and desiccated; drowning, fragmentation and desiccation are all evident for many dammed rivers such as the Paraná²⁸ (Extended Data Fig. 8), Colorado and Mekong rivers. These processes have broad impacts on humans and biodiversity²⁹, though accurate mapping will lead to improved management of water resources and a deeper understanding of connectivity, temporal characteristics and the consequences of land-management decisions³⁰.

The findings reported here reinforce the need for water-resource management strategies that integrate climate and socio-economic dimensions, as has already been proposed¹. This analysis applies a consistent algorithm to all 32 years of the Landsat observations to produce a validated data set that documents global surface water dynamics with new levels of spatial detail and accuracy. Linking this information to complementary data sets, such as satellite altimetry measurements, would produce fresh estimates of surface water volumes, river discharge and even sea-level rise¹². General circulation models that currently treat surface water in a simplistic fashion² should immediately benefit from the accurate location of truly permanent water surfaces. Mapping long-term changes in global surface water occurrence, documenting multi-decadal trends and identifying the timing (to within a given month or year) of events such as lake expansion and retreat or river-channel migration provides insights into the impacts of climate change and climate oscillation on surface water distribution, and concurrently captures the impacts humans have on surface water resource distribution.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 April; accepted 19 October 2016.

Published online 7 December 2016.

- Vörösmarty, C. J., Green, P., Salisbury, J. & Lammers, R. B. Global water resources: vulnerability from climate change and population growth. *Science* **289**, 284–288 (2000).
- Subin, Z. M., Riley, W. J. & Mironov, D. An improved lake model for climate simulations: model structure, evaluation, and sensitivity analyses in CESM1. *J. Adv. Model. Earth Syst.* **4**, M02001 (2012).
- Holgerson, M. A. & Raymond, P. A. Large contribution to inland water CO₂ and CH₄ emissions from very small ponds. *Nat. Geosci.* **9**, 222–226 (2016).
- Gardner, R. C. et al. *State of the World's Wetlands and Their Services to People: A Compilation of Recent Analyses*. Ramsar Briefing Note No. 7, <http://dx.doi.org/10.2139/ssrn.2589447> (Ramsar Convention Secretariat, SSRN, 2015).
- Vörösmarty, C. J. et al. in *Millennium Ecosystem Assessment Vol. 1 Ecosystems and Human Well-being: Current State and Trends* Ch. 7, 165–207, <http://www.unep.org/maweb/documents/document.276.aspx.pdf> (Island Press, 2005).
- World Economic Forum. *The Global Risks Report 2016* 11th edn, <http://www3.weforum.org/docs/Media/TheGlobalRisksReport2016.pdf> (World Economic Forum, 2016).
- Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol.* **296**, 1–22 (2004).
- Downing, J. A. et al. The global abundance and size distribution of lakes, ponds, and impoundments. *Limnol. Oceanogr.* **51**, 2388–2397 (2006).
- Verpoorter, C., Kutser, T., Seekell, D. A. & Tranvik, L. J. A global inventory of lakes based on high-resolution satellite imagery. *Geophys. Res. Lett.* **41**, 6396–6402 (2014).
- Feng, M., Sexton, J. O., Channan, S. & Townshend, J. R. A global, high-resolution (30-m) inland water body dataset for 2000: first results of a topographic-spectral classification algorithm. *Int. J. Digit. Earth* **9**, 113–133 (2015).
- Yamazaki, D., Trigg, M. A. & Ikeshima, D. Development of a global ~90m water body map using multi-temporal Landsat images. *Remote Sens. Environ.* **171**, 337–351 (2015).
- Prigent, C. et al. Changes in land surface water dynamics since the 1990s and relation to population pressure. *Geophys. Res. Lett.* **39**, L08403 (2012).
- Wulder, M. A. et al. The global Landsat archive: status, consolidation, and direction. *Remote Sens. Environ.* **185**, 271–283 (2016).
- Lutz, A. F., Immerzeel, W. W., Shrestha, A. B. & Bierkens, M. F. P. Consistent increase in High Asia's runoff due to increasing glacier melt and precipitation. *Nat. Clim. Chang.* **4**, 587–592 (2014).
- Micklin, P. The future Aral Sea: hope and despair. *Environ. Earth Sci.* **75**, 844 (2016).
- Zafarnejad, F. The contribution of dams to Iran's desertification. *Int. J. Environ. Stud.* **66**, 327–341 (2009).
- van Dijk, A. I. et al. The Millennium Drought in southeast Australia (2001–2009): natural and human causes and implications for water resources, ecosystems, economy, and society. *Wat. Resour. Res.* **49**, 1040–1057 (2013).
- MacDonald, G. M. Water, climate change, and sustainability in the southwest. *Proc. Natl Acad. Sci. USA* **107**, 21256–21262 (2010).
- Mueller, N. et al. Water observations from space: mapping surface water from 25 years of Landsat imagery across Australia. *Remote Sens. Environ.* **174**, 341–352 (2016).
- Tulbure, M. G., Broich, M., Stehman, S. V. & Kommareddy, A. Surface water extent dynamics from three decades of seasonally continuous Landsat time series at subcontinental scale in a semi-arid region. *Remote Sens. Environ.* **178**, 142–157 (2016).
- Postel, S. L., Daily, G. C. & Ehrlich, P. R. Human appropriation of renewable fresh water. *Science* **271**, 785–788 (1996).
- United Nations Department of Economic and Social Affairs, Population Division. *World Population Prospects: The 2015 Revision, Key Findings and Advance Tables*. Working Paper No. ESA/P/WP.241, https://esa.un.org/unpd/wpp/publications/files/key_findings_wpp_2015.pdf (United Nations, 2015).
- Najafi, A. & Vatanfada, J. Environmental challenges in trans-boundary waters, case study: Hamoon Hirmand Wetland (Iran and Afghanistan). *Int. J. Wat. Resour. Arid Environ.* **1**, 16–24 (2011).
- International Commission on Large Dams World Register* http://www.icold-cigb.org/GB/World_register/general_synthesis.asp?IDA=206 (IGIB/ICOLD, 2016).
- Kosarev, A. N., Kostianoy, A. G. & Zonn, I. S. Kara-Bogaz-Gol Bay: physical and chemical evolution. *Aquat. Geochem.* **15**, 223–236 (2009).
- Liu, B. et al. Outburst flooding of the moraine-dammed Zhuonai Lake on Tibetan plateau: causes and impacts. *IEEE Geosci. Remote Sens. Lett.* **13**, 570–574 (2016).
- Pethick, J. & Orford, J. D. Rapid rise in effective sea-level in southwest Bangladesh: its causes and contemporary rates. *Glob. Planet. Change* **111**, 237–245 (2013).
- Bonetto, A. A., Wais, J. R. & Castello, H. P. The increasing damming of the Paraná basin and its effects on the lower reaches. *Regul. Rivers Res. Manage.* **4**, 333–346 (1989).
- Vörösmarty, C. J. et al. Global threats to human water security and river biodiversity. *Nature* **467**, 555–561 (2010).
- Acuña, V. et al. Why should we care about temporary waterways? *Science* **343**, 1080–1081 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements The USGS and NASA provided the Landsat imagery. R. Moore and her team provided the Google Earth Engine. R. Sargent and P. Dille from Carnegie Mellon University built the web interface to the global surface water occurrence maps, and M. Clerici and J. van 't Klooster built the web processing interface.

Author Contributions Each author contributed extensively and indispensably to the work presented in this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.-F.P. (jean-francois.pekel@jrc.ec.europa.eu).

Reviewer Information *Nature* thanks I. Klein and D. Yamazaki for their contribution to the peer review of this work.

METHODS

Data. The entire archive of the Landsat 5 Thematic Mapper (TM), the Landsat 7 Enhanced Thematic Mapper-plus (ETM+) and the Landsat 8 Operational Land Imager (OLI) orthorectified, top-of-atmosphere reflectance and brightness temperature images (L1T)³¹ acquired between 16 March 1984 and 10 October 2015 was used^{32,33}.

Landsats 5, 7 and 8 are in near-polar orbit with repeat cover every 16 days; two satellites operate concurrently with an 8-day cycle. The ground area imaged by adjacent orbits overlaps by 7.3% at the Equator, increasing to 68.3% overlap at 70° latitude³⁴. Adjacent orbits to the west of a previous one are collected a week apart, thus pixels in the side-lap areas from both orbits are unique views. Sequential image frames along an orbit track also overlap, but these are identical data, not unique views; we exclude the end-lap but retain side-lap pixels for analysis.

Landsat 5, launched 1 March 1984, collected TM imagery until November 2011. Landsat 7 was launched 15 April 1999 and acquired imagery normally until 31 May 2003 when the scan line corrector (SLC) failed³⁵, and Landsat 8 began operational imaging in April 2013. The SLC failure causes around 22% of each scene to be lost³⁶. This loss increases from the centre, giving SLC-off images a slatted appearance at the edges.

Extended Data Fig. 1a–c shows variations in geographic coverage, first year of imaging and the number of images acquired from year to year. Landsat 5 had no on-board data recorders and its links with data relay satellites failed over time, so cover was often limited to the line of sight of receiving stations, and these did not provide complete global coverage³⁷. Commercial management of the programme from 1985 to the early 1990s³⁸ meant that acquisitions tended to be acquired only when pre-ordered³⁵. Geographic and temporal unevenness are particularly apparent up to 1999; the Americas, Western Europe and Africa were first imaged in 1984, Australia and South East Asia in 1986, but Kolyma was not imaged until 1995 and the northeastern part of the Siberian plateau not until 1999. These gaps are a feature of all global data sets based on the Landsat 30-m archives³⁹. Two Landsats operated from April 1999 onwards, which improved coverage, especially as Landsat 7 had on-board data recording capabilities. The programme's Long-Term Acquisition Plan also began in 1999, which further improved global coverage⁴⁰. Acquisitions fell between 2011 and 2012 as Landsat 5 operations were curtailed, though they increased in 2013 when Landsat 8 joined Landsat 7. More images were collected in June, July and August than December, January and February in all years, except for June 2003, following the SLC failure. However, by August 2003 data were once again routinely acquired (albeit with SLC-off gaps). From April 2013 the programme collected almost all images over land and in 2010 the USGS also began their Landsat Global Archive Consolidation, which ingests historic archive holdings from receiving stations around the world and produces orthorectified, top-of-atmosphere (L1T) data from these¹³. At the time of this study 3,066,080 L1T images (1,823 terabytes of data) covered 99.95% of the landmass. These images provided a local maximum number of unique views (including the side-lap areas) of 1881 and minimum of 11; the global mean was 537 and median 482. Pixels were excluded where over- or under-saturation occurred (manifested as random speckle or line-dropout), where one or more spectral bands were missing (typically occurring at image edges), and where scene geo-location was compromised.

Water detection. Although the goal of this study may seem simple—to separate water from other surfaces—consistent discrimination on the global scale, over multiple decades, is a non-trivial challenge. Water is a highly variable target and its spectral properties at the wavelengths measured by the TM, ETM+ and OLI sensors vary according to chlorophyll concentration, total suspended solids and coloured dissolved organic matter load, depths and water-body bottom material for shallow waters⁴¹, as well as variations in observation conditions (sun-target-sensor geometry, and optical thickness). On the global scale over 32 years all these conditions will be encountered somewhere at some time.

To address the challenge outlined above, techniques for big data exploration and information extraction less commonly used by the remote sensing community were exploited, namely expert systems^{42–44}, visual analytics⁴⁵ and evidential reasoning⁴⁶. Expert systems provide flexibility (to deal with the range of conditions encountered), visual analytics combine human cognitive and perceptual abilities with the storage and processing capacities of cloud computing platforms (the overall premise being that experts will have a better understanding of the problem to be solved if they interact with the data and view it through different representations and are thus able to design, test and fine-tune scenarios to extract the desired information), and evidential reasoning deals with problems related to both uncertainties and quality issues in the data set.

Expert systems are non-parametric classifiers that can account for uncertainty in data, incorporate image interpretation expertise into the classification process, and can be used with multiple data sources. The expert system outlined in Extended Data Fig. 2 was developed to assign each pixel to one of three target classes, either

water, land or non-valid observations (snow, ice, cloud or sensor-related issues). The inference engine of our system was a procedural sequential decision tree, which used both the multispectral and multitemporal attributes of the Landsat archive as well as ancillary data layers. Within the inference engine, expert knowledge was represented in the form of rules having the form: IF condition THEN inference. The condition contains equations describing the cluster hulls in a defined multispectral feature-space and can also be a combination of logical statements in which several components are linked through logical operators. The chaining of IF–THEN rules forms the problem-solving model that organizes and controls the steps and data used in the classification. Tracing the line of reasoning used by the inference engine during the development phase meant that the reason for the class choice associated with each pixel could be retrieved, which in turn allowed the reasons behind particular classification challenges to be identified. Solutions to identified failings could then be developed using evidential reasoning and addressed in subsequent iterations so that the overall performance of the classifier was progressively improved. When the performance of the expert system could no longer be noticeably improved it was applied to the entire Landsat L1T data set in a single run, and the output of this classification was then validated.

The equations describing the cluster hulls used in the expert system were established through visual analytics. The first step was to build a spectral library capturing the spectral behaviour of the three target classes across as wide a range of conditions as possible. 64,254 samples obtained through visual interpretation of 9,149 Landsat scenes recorded spectral variability of the target classes. Records held in the library comprised spectral values from all bands. These records were enriched by deriving the Normalized Difference Vegetation Index and Hue–Saturation–Value (HSV) colour-space transformations for the following band combinations: shortwave infrared (SWIR2); near-infrared (NIR); red; and NIR/green/blue using a standard transformation⁴⁷. The HSV colour model is well adapted for image analysis because the chromaticity (H and S) and the overall brightness (V) components are decoupled. This is highly desirable because changes in observation conditions first affect the V component and then the S component, while H remains relatively stable (except when the fundamental nature of the target changes, such as when land becomes water). Consequently, this property promotes temporal stability in the measurements and HSV-based classifications have been successfully used for near-real-time surface water detection at continental scales⁴⁸.

The information held in the spectral library was then analysed through visual analytics with the goal of extracting equations describing class cluster hulls in multidimensional feature-space for the expert system. An exact representation of clusters in feature-space helps to visualize class spectral overlap and thus to provide an understanding of the sources of thematic class uncertainty. Two-dimensional representations of the feature-space occupied by the samples held in the spectral library are shown in Extended Data Fig. 3. An exploratory data analysis tool was designed to support the interactive analysis. This provided multiple coordinated views⁴⁹ of the scatter plots and the source image from which any particular point within the plots was obtained. This two-way link between the representations of the samples within the multidimensional feature-space and source imagery meant that the geographic location (and context) of all samples could be considered along with their behaviour in feature-space. This was particularly valuable where the spectral properties of water coincided with those of other target classes. In these cases all dimensions of the feature-space could be examined to determine whether spectral separation was indeed possible.

The expert supervising the inference engine's development could then interactively draw the vertices of the hulls in this feature-space, which were then converted into equations through Delaunay triangulation⁵⁰ (Supplementary Table 2). Drawing the hulls directly into feature-space allowed the expert to account for irregular (even concave) shapes of clusters associated with the three target classes, and to deal adequately with the skewed data distributions so often encountered in remote sensing data sets. This approach also allowed cluster hulls to be defined that captured very infrequent—but nevertheless thematically important spectral behaviour. For example, the heavily sediment-laden seasonally occurring waters in parts of West Africa do not often appear in the time series, and are therefore poorly represented in the scatter plots, but these rare occurrences are important. The resulting equations are perforce complex because they constitute a precise representation of the complex shape of the three clusters within the multidimensional feature-space.

However, not all pixels could be unambiguously assigned to one or other of the target classes because overlap between the clusters was such that it could not be resolved anywhere in multidimensional feature-space. In these instances evidential reasoning—as part of the expert system—was used to guide class assignment. This took into consideration geographic location and the temporal trajectory in multispectral feature-space. Geographic location is important because spectral confusion between water and other surfaces can only occur in locations where

those other surfaces are found, and time is important because this spectral overlap may occur at specific times of the year and not at others.

The temporal trajectory in feature-space was used to gain information concerning the likelihood of a pixel being water. If a pixel sits unequivocally within a water hull for some of the time, then there is a high likelihood it will actually be water even if it occasionally occupies a hull where overlap occurs with other cover types, whereas if a pixel always sits in the overlapping domain or moves to the hull delimiting land this likelihood is removed. The frequency with which a pixel occupies the unequivocal portion of the water hull is used to estimate the likelihood of it actually being water.

Geographic locations linked to specific sources of spectral overlap were delineated using ancillary data layers. These constrain a portion of the data set within which specific decision rules could be identified and applied.

To detect water over glaciers spectral overlap from supraglacial moraines and shadow had to be resolved. The glacier areas were first delineated using the Randolph Glacier Inventory 5.0⁵¹. This defined the geographic regions where specific decision rules were required to confirm the water class assignment. Actual water over glaciers (melt ponds and surface streams) sometimes occupied the unequivocal portion of the water cluster in feature-space and the frequency with which this occurred determined the likelihood of water presence at that location.

Lava flow is also often wrongly assigned to the water class. Part of the spectral overlap between lava and water occurs in all dimensions of the multispectral feature-space, although again, over time, a pixel may move into a part of the feature-space where there is no overlap. A global-scale lava mask was established from both spectral characterization and visual interpretation of Landsat images, and within these boundaries the frequency with which pixels occupied 'lava overlap-free' portions of the water cluster were computed over the full span of the archive. This frequency was again used as an indicator of the likelihood of water presence for those locations.

Shadow, from whatever source, can cause false water detections because the underlying spectral characteristics of the surface are not truly represented, and may overlap with water. Three sources of shadow were addressed; buildings, terrain and clouds.

In the case of buildings, location and spectral behaviour over time were again brought into play. The Global Human Settlement Data Layer (GHSL)^{52,53} targeted areas where spectral confusion between water and the shadows cast by buildings needed to be resolved. Building shadows are seasonally dependent (related to changes in solar zenith angle), and thus a pixel may move in and out of the 'shadow-water overlap' cluster over time. This movement may be to a land cluster, or water. Over time, if a pixel exclusively moves to the water cluster, that location is very likely to be a permanent water feature within the urban area. The higher the frequency with which a pixel occupies the unequivocal water hull the greater the probability of water's presence at that location. This allows the expert system to map permanent water features in urban areas, and also to account for loss of water due to urban expansion and land reclamation. However, seasonal water detections within urban areas are more problematic, because these pixels will move between land and water in multispectral feature-space over time even in the absence of shadow.

Terrain shadows may be resolved using a threshold applied on slopes derived from a Digital Elevation Model (DEM). Unfortunately, this is only valid if the derived slopes represent the conditions on the date a given Landsat scene is acquired. Reservoirs are often built in steep terrain, and any dam built after the release date of the DEM would be masked, because the slopes recorded in the DEM are those prevailing before the reservoir filled. To detect new reservoirs, a mask covering areas where terrain shadow is expected was first derived from one of four DEMs^{54–57} (four were used to provide the best available DEM resolution at any given location). Within the masked area, pixels detected as water at multiple dates across the full year are likely to be water rather than seasonally cast shadow. Locations where water is detected in months when the sun is close to nadir are especially likely to be water. And again the movement into and out of the unequivocal water cluster in feature-space over time was used to identify actual water.

Cloud shadow is an even greater challenge than that from terrain or buildings, because it can occur anywhere and anytime. Thus the cloud shadow mask needed to be produced dynamically. Established cloud detection routines, such as FMASK⁵⁸, provide scene-by-scene identification of the location and extent of cloud shadow as this is linked to the clouds they detect. This supposes that the cloud producing the shadow is present in the image. But clouds outside a scene may still cast shadows in adjacent scenes. However, images from adjacent orbits cannot be used to resolve this because these occur a week apart. Consequently, cloud detection cannot be used as a robust indicator of cloud shadow. However, cloud shadows move over time, while water surfaces show more consistent temporal behaviour (even seasonal water can be expected in some months, and not at all in others). A temporal sliding

window was used across the water history record, post-classification. Within this sliding window, if the preceding and subsequent values were identified as water, then the observation under consideration was also likely to be water. Water detections bracketed by land detections are much more likely to be shadows. At these locations the frequency with which pixels occupied the unequivocal portion of the water cluster was again considered across the time series, and the lower this frequency the greater the likelihood of shadow.

Finally, visual inspection of the water maps identified scattered residual false detections, which were manually removed. Less than 0.002% (72 km²) of the maximum water extent was cleaned in this way. These errors were linked to industrial sites, photovoltaic farms and urban infrastructure not represented in the GHSL^{52,53}, such as airport runways. Some such errors may still remain, but their impact is accounted for in the validation.

While the drawing of the cluster hulls and the evidential reasoning are subjective, the visual analytics guide and inform the expert's decisions, which are based on objective use of the spectral library and the associated contextual information held in the image archive. Thus, this subjectivity is compensated for by an increased degree of confidence in the analysis⁵⁹ and the absence of biases associated with *a priori* assumptions concerning normal distributions associated with supervised classifiers such as the maximum likelihood or Mahalanobis distance.

The expert system was run in Google's Earth Engine, a computational infrastructure optimized for parallel processing of geospatial data plus a dedicated application programming interface and online access to the USGS L1T Landsat archive. Running the expert system on a single computer central processing unit (CPU) would have taken 1,212 years, but using 10,000 computers in Earth Engine the processing was completed in around 45 days, although building, testing and validating the expert system took almost two years.

Code availability. A web interface powered by Google Earth Engine allows the expert system to be run on any Landsat 5, 7 and 8 images. Access can be provided on request.

Validation. The expert system's performance was judged in term of errors of omission and commission at the pixel scale. The validation design took into account the small spatial extent of inland water surfaces (3% of the land surface) and its intrinsic spatio-temporal variation. The validation was performed using a total of 40,124 control points distributed both geographically (globally), temporally (across the 32 years), and across sensors (TM, ETM+ and OLI). Extended Data Fig. 4 summarizes the validation protocol.

Two reference data sets were produced: a sample of 27,268 pixels was dedicated to the estimation of the error of omission and 12,856 pixels were used to characterize errors of commission. Extended Data Fig. 5 shows the geographic distribution of all sample points and the associated errors.

To generate the omission error data set, a systematic sample frame (a grid 1° latitude by 1° longitude) was used. Under this frame the globe was stratified into areas with a high probability of water occurrence based on existing published global surface water maps, one for all latitudes up to 60° N (ref. 60) and one from 60° N to 78° N (ref. 10). A point was randomly selected within the water reference stratum in each latitude/longitude grid cell, and an image with less than 10% cloud cover corresponding to this location was then randomly selected for each sensor across the time span of the archive. In some locations the target of one sample per sensor at each location was not reached because of the absence of water at certain times, the total absence of observations over part of the archive or frequent cloud coverage.

Both of the published global water maps represent specific time periods and cannot guarantee the presence of water across all seasons and at all dates. The presence of water for each single validation point at each date was confirmed by visually checking all points using the validation tool described below. Only points confirmed as water were used in the estimation of the omission error.

To determine commission errors a grid 1° latitude by 1° longitude was again used as a systematic sample frame. For each latitude/longitude grid cell, images were randomly selected from the archives for all three sensors and the water detection expert system was run on these. One point per sensor from the areas classified as water was then randomly selected. The actual presence of water for each single validation point at each date was visually checked, again using the validation tool. If no water was in fact present, then this corresponded to an error of commission.

The validation tool was designed to facilitate the photointerpretation task. For each pixel in the database the full-resolution Landsat image selected for the validation (that is, randomly allocated in time to those pixel coordinates) was displayed and the specific pixel for interpretation highlighted. Images from the same location acquired at dates before and after that of the sample image were also presented, along with high-resolution satellite imagery and/or aerial photography from Google Earth and the Environmental Systems Research Institute (ESRI).

Using all these inputs an expert interpreter confirmed or refuted the presence of water at that location and time. The database was automatically populated with the results.

Errors of omission overall were less than 5% and commission less than 1%. Breaking down the validation sample by sensor showed that all three performed comparably well; commission errors ranged by only 0.2% and omission by 1.2%. Extended Data Table 1 provides details.

Breaking down the validation sample further by water seasonality class (using the seasonality class for the relevant year in which the validation point was randomly acquired) shows that although all three sensors performed similarly, omission for seasonal water classes was, as expected, higher than that for permanent water classes. Accuracy for TM, ETM+ and OLI when judged against errors of commission were respectively 99.6%, 99.5% and 99.7%, for permanent water, while accuracies considering errors of omission for permanent water were 98.8%, 97.8% and 99.1%. Errors of commission for seasonal water were very slightly higher, with accuracies of 98.8% (TM), 98.4% (ETM+) and 98.5% (OLI) and the greater errors of omission for seasonal water were reflected in accuracies of 74.9% (TM), 73.8% (ETM+) and 77.4% (OLI).

Errors of omission for seasonal water are higher than for permanent because there are fewer opportunities to observe each water body. These errors will result in an underestimation of its occurrence. But omission errors do not occur in the same place over time and over the 32 years multiple opportunities for observation arise. Thus sites where seasonal water can occur may be missed at one date, but may be correctly mapped at another. Overall, less than 1% of the points (214 samples out of 27,268) in the validation database where water was actually present remained entirely unmapped over time (Extended Data Table 1).

Thematic mapping. The maximum water extent (all locations ever mapped as water) during the 32 years was mapped. The frequency with which water was present on the surface from March 1984 to October 2015 was captured in a single product called surface water occurrence (SWO). To compute SWO, the water detections (WD) and valid observations (VO) from the same months are summed, that is, water detections and valid observations from March 1984 are added to water detections and valid observations from March 1985 and so on, such that $SWO_{\text{month}} = \sum WD_{\text{month}} / \sum VO_{\text{month}}$. Averaging the results of all monthly SWO_{month} calculations gives the long-term overall surface water occurrence. The month-by-month time step normalizes occurrence against seasonal variation in the number of valid observations across the year. Typically, more cloud-free observations (and thus valid observations) are available during dry seasons than wet. Without monthly weighting, the overall water occurrence (that is, computed over the full period) would be biased by temporal distribution of the valid observations (that is, giving more weight to the dry season than to the wet season). Extended Data Fig. 1c documents the number of scenes per month and year across the archive.

Change in water occurrence intensity between two epochs (16 March 1984 to 31 December 1999, and 1 January 2000 to 10 October 2015) was also produced (Extended Data Fig. 6a). This is derived from homologous pairs of months (that is, the same months contain valid observations in both epochs). The occurrence difference between epochs was computed for each pair and differences between all homologous pairs of months were then averaged to create the surface water occurrence change intensity map. Areas where there are no pairs of homologous months could not be mapped. The averaging of the monthly processing mitigates variations in data distribution over time (that is, both seasonal variation in the distribution of valid observations, temporal depth and frequency of observations through the archive) and provides a consistent estimation of the water occurrence change. This map shows where surface water occurrence increased, decreased or remained invariant between the two epochs.

The occurrence and occurrence change intensity maps provide a summary of the location and persistence of water on the surface, but they do not describe inter- and intra-annual variability. We propose water recurrence as a measurement of the degree of inter-annual variability in the presence of water. This describes how frequently water returned from one year to another (expressed as a percentage). Recurrence refers specifically to the temporal behaviour of water surfaces; unlike occurrence, recurrence is not systematically computed over the full span of the archive, because water may not have been present from the beginning to the end of the archive. Thus, we first have to define a 'water period'—that is, that part of the archive where water was present at least from time to time; the recurrence in fact quantifies this 'time to time'. The water period is established individually for each pixel. The water period runs from the first month in the first year in which water is observed to the last month of the last year in which water is observed of the entire 32-year period. In addition to defining the water period we also need to define a 'water season' (not equivalent to a 'wet season'). The water season is identified from the monthly water recurrence and is defined as those months of the year that from

time to time have water. A 'water year' is a year with at least one water observation, while an 'observation year' is a year with at least one valid observation within the water season. Water recurrence is then calculated as the ratio of the number of water years to observation years. The count of the number of years starts with the year in which water was first observed and ends with the most recent year in which water was observed. Years that contain only observations outside the water season are not counted; we have no way of knowing whether water might have occurred in the water season because we have no observations.

We also describe the intra-annual distribution of water, which discriminates between 'permanent' and 'seasonal' water surfaces. A permanent water surface is underwater throughout the year, while a seasonal water surface is underwater for less than 12 months of the year. In some places we do not have observations for all 12 months of the year (for example, because of the polar night in winter) and in these cases water is considered to be seasonal if the number of months where water is present is less than the number of months where valid observations were acquired. A second consideration is that some lakes freeze for part of the year. However, during the frozen period water is still present under the ice layer, both for lakes and the sea. The expert system treats ice as a non-valid observation, so the observation period corresponds only to the unfrozen months. If water is present throughout the observation period (that is, the unfrozen period), the lake is considered to be a permanent water surface. If the area of the lake contracts during the unfrozen period, then the pixels along the borders of the lake no longer represent water, and those pixels will be considered to represent seasonal water surface.

Seasonality is computed for every year. A single data set for the contemporary period (October 2014 to October 2015) is made available via the website and Extended Data Fig. 7a provides an example. The individual years' computations are used to produce the trends analyses provided in Fig. 3. Plotting the measured permanent surface water area for each year would provide such trends. But the gaps in the observation record are a source of uncertainty, especially in the early years of the archive. Consequently, part of the permanent water surface is potentially not taken into account (not observed), and constitutes a source of underestimation of the reported area. To account for this we combine the measured values of permanent surface water area with an estimate of the area of unobserved but potentially permanent surface water. This is computed using the maximum permanent water extent, that is, any pixel that has ever been identified as permanent in any year of the record across the full 32 years, minus the observed land values, minus any observed seasonal water, minus the observed permanent water. The true permanent surface water area will lie somewhere within this unobserved range, but the actual limits cannot be established. There is no uncertainty in those instances where the observation record is complete; conversely, in those years where no observations were made uncertainty is absolute.

Trends are calculated for defined geographic regions. The derived trend parameters (water loss per year, correlation, *P*-value) were computed using years where this unobserved area was less than 5% of the observed area. For countries and state-level reporting the Global Administrative Areas dataset (GADM)⁶¹ was used, for the Tibetan Plateau the boundary established by the Chinese Institute of Geographic Sciences and Natural Resources Research⁶² was used, and the trends over the Aral Sea were delimited by top-left coordinates 47° N, 58° E, bottom-right 43.8° N, 61.5° E.

Temporal profiles. Three histograms are generated. First, monthly recurrence shows the intra-annual distribution of the water, and characterizes water seasonality. It also provides information on the water recurrence for each single month. Second, a water history chart shows the class (land, seasonal water and permanent water) for each year in which valid observations were acquired. Third, month-by-month presence of water and observations within any single year can be extracted. Examples can be seen in Extended Data Fig. 6b.

Measuring change. The thematic maps and temporal profiles were used to identify a set of water classes that characterize transitions between the first year in which representative observations were acquired and the last year of observation. Representative years are identified by comparing each year in turn with the annual pattern of monthly recurrence from the temporal profiles. These profiles identify months in which water was observed, and indicate the percentage of valid observations classified as water in any given month. A year is flagged as representative if it contains sufficient valid observations from any combination of months to bring confidence to the determination of the presence or absence of water. The overall level of confidence is determined by the annual sum of the monthly long-term recurrences of observed months (per year). The rationale is that the likelihood of a real absence of water for a year is higher if the water is absent for months showing a high long-term water recurrence than from one showing small rates of recurrence. In the latter case the absence of water may be explained by a seasonal shift, and does not confer enough confidence to conclude that water was not present later.

Therefore, we considered that if the sum of the recurrence of the observed months is greater than 100, the absence of water observation brings enough confidence to consider that water was actually not present. Conversely, a single water presence is enough to demonstrate water presence. The water class in that representative year is then fixed as the 'first' year. The last year's water class is always the class assigned to the last year of observation (October 2014 to October 2015) because we have enough observation available within a year during this period.

The following transitions were mapped: unchanging permanent water surfaces; new permanent water surfaces (conversion of land into permanent water); lost permanent water surfaces (conversion of permanent water into land); unchanging seasonal water surfaces; new seasonal water surfaces (conversion of land into seasonal water); lost seasonal water surfaces (conversion of a seasonal water into land); conversion of permanent water into seasonal water; and the conversion of seasonal water into permanent water (Extended Data Fig. 8 provides an example).

These conversions refer to changes in state from the beginning and end of the time series; they do not describe what happened in the intervening years, so an unchanging water surface means that the seasonality at that particular point was the same in the first and last year it was observed, and not necessarily that it was stable throughout. Stability must be checked at the pixel scale by using the long-term water history described by the temporal profiles plus the recurrence and occurrence maps. There are instances where water is not present at the beginning or the end of the observation record but is present in some of the intervening years. By tracking the inter-annual patterns of such 'ephemeral' events and their intra-annual characteristics, each such pixel can be classified as either ephemeral permanent water (land replaced by permanent water that subsequently disappears) or ephemeral seasonal water (land replaced by seasonal water that subsequently disappears), depending on the majority of the observed seasonality during the period of water presence.

The GADM⁶¹ was also used to extract water area statistics at national, continental and global levels from the occurrence, recurrence and transition maps plus the long-term water history. Area measurements (in km²) for the following classes are reported in Supplementary Table 1: maximum surface water occurrence over 32 years; permanent water in the first year of observation; permanent water in the last year of observation; permanent water with 100% recurrence; transition from land to permanent water; transition from seasonal to permanent water; transition from permanent water to land; transition of permanent water to seasonal; seasonal water in the first year of observation; seasonal water in the last year of observation; and seasonal water with 100% recurrence.

Known issues and planned improvements. Bodies of water smaller than 30 m by 30 m, those obscured by floating, overhanging and standing vegetation or hidden by infrastructure such as tunnels and bridges were not included. Irrigated fields that stand in open water for some weeks were mapped but not when crop cover is well established. When observations coincide with paddy flooding, yet predate crop emergence and cover, then paddy fields are mapped, but inevitably some will be missed. Paddy fields in steep terrain present particular challenges because of their small size. Paddy fields are an example of short-duration events, but in fact short-duration seasonal water more generally is likely to be underestimated because of geographic and temporal discontinuities in the archive and gaps caused by persistent cloud cover.

The precision of the metrics at any location improves as the number of valid observations increases. The meta-information in the web interface documents the number of valid observations at each pixel location, which provides users with a proxy measure of confidence.

Long-term changes cannot be determined uniformly for the entire globe because the observation record varies; the first year of observation is 1984 for much of the world, but not for parts of the Siberian plateau and Kolyma (1999 and 1995 respectively). For these reasons we recommend caution in interpreting the changes in these particular regions, though given the geographic completeness of current L1T coverage we have confidence in the contemporary figures reported. The water detections are accurate (as determined by the validation), but the time range over which the transitions occurred in these regions is perforce shorter. Northern Hemisphere high latitudes are also problematic because the observation season is short, the solar zenith angles are low, and the archive is poorly populated above 78° N. One consequence is that at high latitudes, the seasonality of the ocean is occasionally wrongly reported as seasonal.

On the surface of glaciers, melt ponds and streams are identified here as permanent water surfaces, but this permanence is questionable, because unlike deeper lakes where only the surface freezes, here the full water volume may actually freeze.

The recurrence value may be overestimated for water surfaces that came into existence in the last few years of the time series. The temporal profiles identify

the years in which water occurred, and using these, decisions can be taken as to whether recurrence is considered relevant at any given pixel location.

Some isolated commission errors still occur in urban areas as the GHSL^{52,53} is itself being improved; roofs, coal and waste heaps and runways are the most common sources of confusion. These will be resolved as the urban information layers improve.

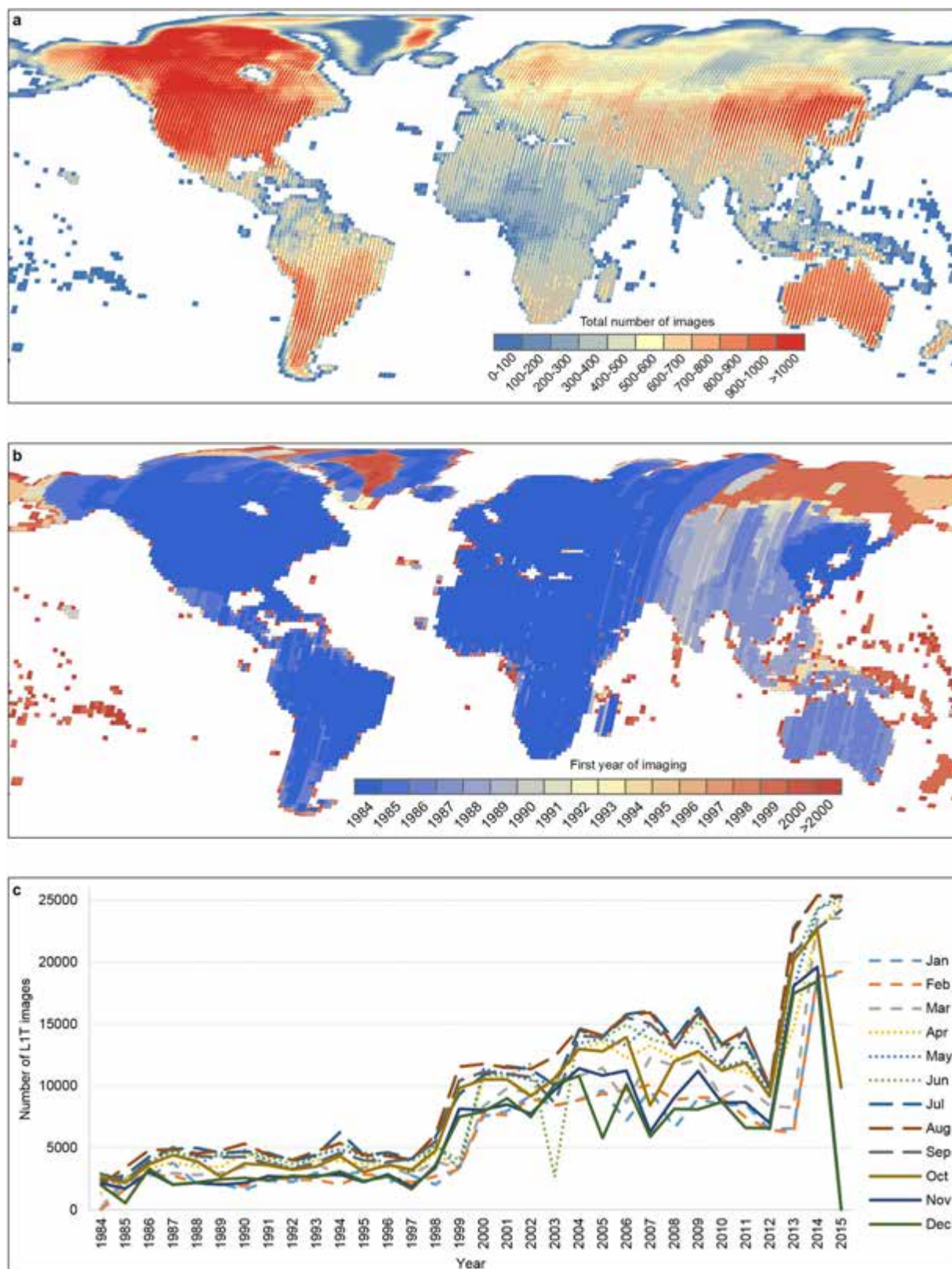
The SLC-off condition also introduces artefacts because the slatted appearance of the original images is occasionally carried into the water occurrence maps. These are not misclassified pixels but occur because the sampling between the gaps is greater than within them, allowing different water conditions to be captured. Techniques to fill the SLC-off gaps exist³⁶, but these create artificial values, and considering the strong temporal and spatial variability of water these techniques require careful use, as they may create false water detections. Water mapping using multitemporal time series on continental scales avoided the use of such techniques on these grounds¹⁹, though did note the same issue of SLC-off artefacts appearing in the final maps.

Analysis of coarse-spatial-resolution satellite data sets from systems offering daily revisit (nearly one observation a day over each location) first captured the inter- and intra-annual variability of surface water occurrence^{12,63,64}, and although the Landsat missions offer 8-day or 16-day rather than daily observations, they do provide high-resolution land surface observations spanning more than three decades. The spatial and temporal information reported in this data set complements that acquired in the past. Nevertheless, the biggest limitation to global surface water occurrence mapping from these data are undoubtedly the geographic and temporal discontinuities of the archive itself. The Landsat archive is continuously enriched through new acquisitions and recovery of old data from international receiving stations¹³. Imagery from other satellites in this resolution class could also be used to improve the temporal sampling. At least 24 other satellites have gathered multispectral imagery at resolutions of 20 m to 30 m from near-polar orbits concurrent with the Landsat programme⁶⁵. These are managed by at least 12 different sovereign states, and although data access is not always at the exemplary full, free and open level of Landsat, some systems do provide this, for example the European Union's Sentinel 2a satellite launched in 2015; the next version of the expert system will also ingest these data streams. Landsat 4, which was launched 16 July 1982 and in operation until 14 December 1993 also carried a 30-m-resolution TM, though the satellite contributes only a fraction to the total TM holdings in the archive, and much of this was restricted to the conterminous USA in the first two years of operation³⁶. Nevertheless, future reanalysis will include the Landsat 4 data and could possibly be extended back to 1972 through the inclusion of data from the Landsat Multi Spectral Scanner (MSS), though this would be challenging because of the more limited spectral, spatial and temporal dimensions of these data sets⁶⁶. Combining all available satellite observations with petabyte processing power would put real-time monitoring of change to Earth's inland and coastal waters within reach.

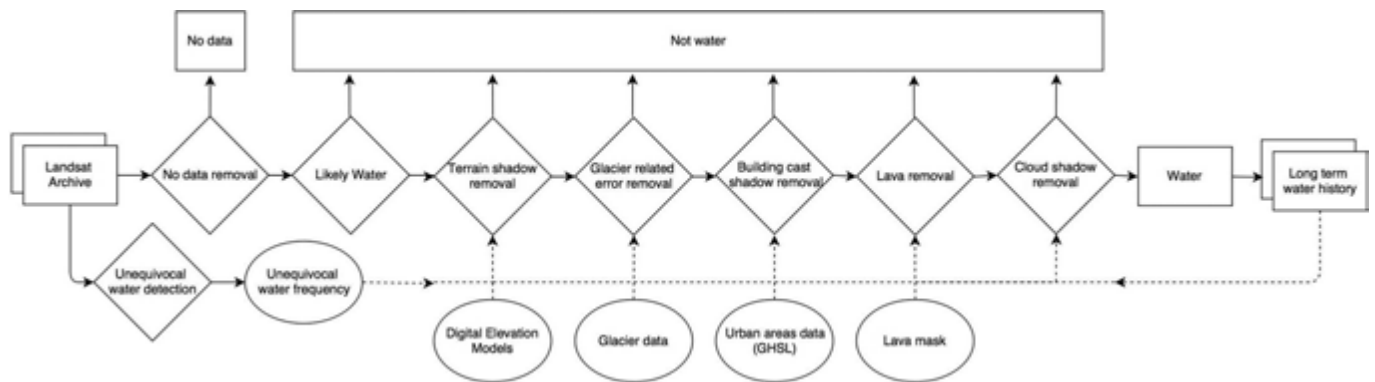
Data availability. The Landsat imagery used in this study is available from the USGS <http://earthexplorer.usgs.gov> and in Google Earth Engine <https://earthengine.google.com>. The data sets generated during the current study are available from <https://global-surface-water.appspot.com>. The data used to generate Fig. 2 and Extended Data Fig. 1c are provided as Source Data.

1. *Landsat 8 Data Users Handbook* <http://landsat.gsfc.nasa.gov/?p=10659>, USGS Publication LSDS-1574 (US Geological Survey, 2016).
2. Woodcock, C. E. *et al.* Free access to Landsat imagery. *Science* **320**, 1011 (2008).
3. Wulder, M. A. *et al.* Opening the archive: how free data has enabled the science and monitoring promise of Landsat. *Remote Sens. Environ.* **122**, 2–10 (2012).
4. *Landsat 7 Science Data Users Handbook* http://landsathandbook.gsfc.nasa.gov/orbit_coverage/prog_sect5_2.html (NASA, accessed 16 November 2016).
5. Markham, B. L., Storey, J. C., Williams, D. L. & Irons, J. R. Landsat sensor performance: history and current status. *IEEE Trans. Geosci. Remote Sens.* **42**, 2691–2694 (2004).
6. Chen, J. *et al.* A simple and effective method for filling gaps in Landsat ETM+ SLC-off images. *Remote Sens. Environ.* **115**, 1053–1064 (2011).
7. Goward, S. *et al.* Historical record of Landsat global coverage. *Photogramm. Eng. Remote Sensing* **72**, 1155–1169 (2006).
8. Loveland, T. R. & Dwyer, J. L. Landsat: building a strong future. *Remote Sens. Environ.* **122**, 22–29 (2012).
9. Gutman, G. *et al.* Assessment of the NASA-USGS global land survey (GLS) datasets. *Remote Sens. Environ.* **134**, 249–265 (2013).
10. Arvidson, T., Gasch, J. & Goward, S. N. Landsat 7's long-term acquisition plan—an innovative approach to building a global imagery archive. *Remote Sens. Environ.* **78**, 13–26 (2001).
11. Arst, H. *Optical Properties and Remote Sensing of Multicomponental Water Bodies* Vol. XII of *Marine Science and Coastal Management* Ch. 1 (Springer Science Praxis, 2003).
12. Lu, D. & Weng, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* **28**, 823–870 (2007).

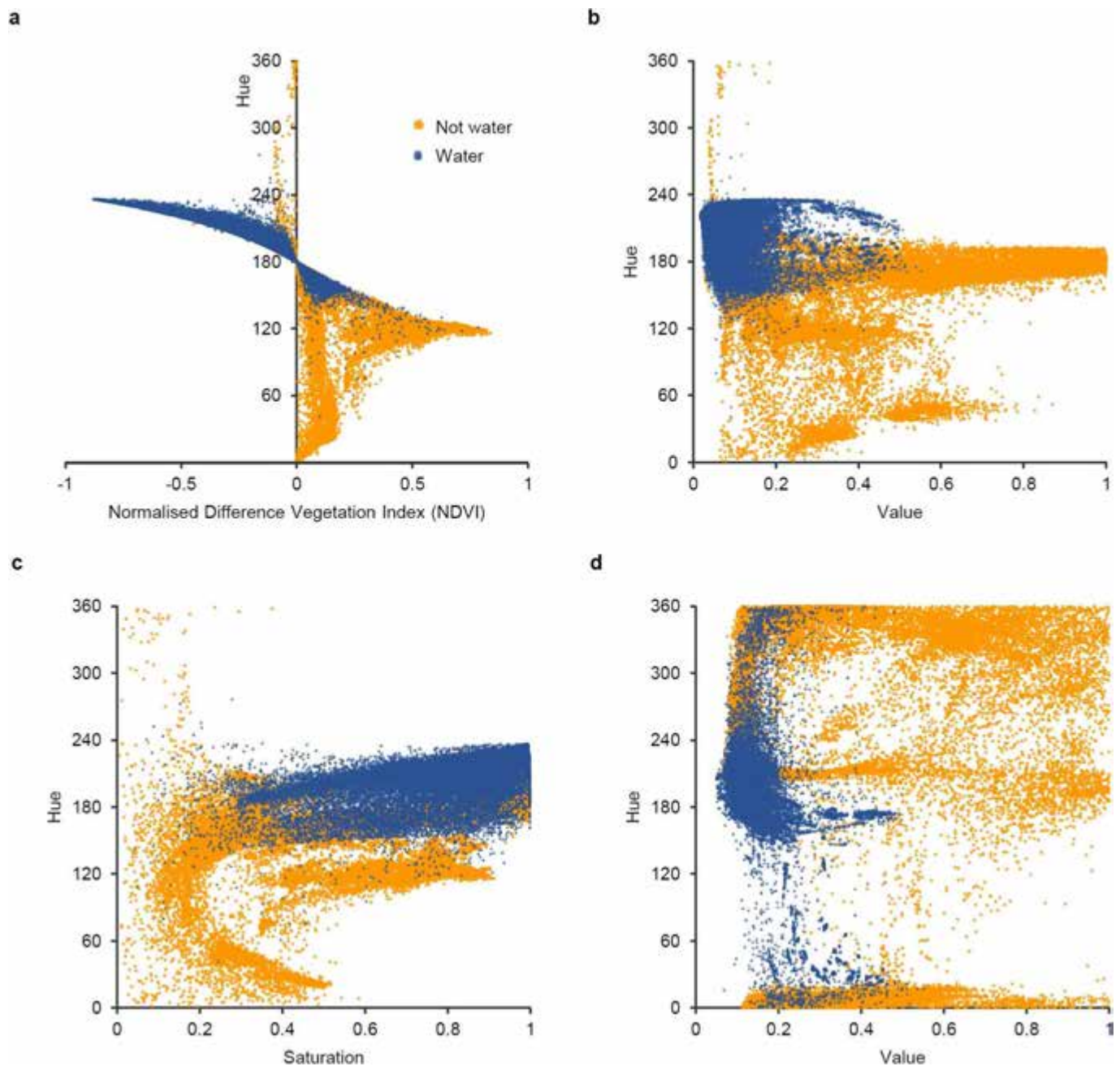
43. Kartikeyan, B., Majumder, K. L. & Dasgupta, A. R. An expert system for land cover classification. *IEEE Trans. Geosci. Remote Sens.* **33**, 58–66 (1995).
44. Shoshany, M. Knowledge based expert systems in remote sensing task: quantifying gains from intelligent inference. *Int. Soc. Photogramm. Remote Sens. Arch.* XXXVII (B7) 1085–1088, http://www.isprs.org/proceedings/XXXVII/congress/7_pdf/6_WG-VII-6/06.pdf (XXIst ISPRS Congress, Technical Commission VII, 2008).
45. Keim, D. A. et al. in *Visual Data Mining* 76–90, http://kops.uni-konstanz.de/bitstream/handle/123456789/5631/Visual_Analytics_Scope_and_Challenges.pdf?sequence=1&isAllowed=y (Springer, 2008).
46. Yang, J.-B. & Xu, D. L. On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Trans. Syst. Man Cybern. A* **32**, 289–304 (2002).
47. Smith, A. R. Color gamut transform pairs. *Comput. Graph.* **12**, 12–19 (1978).
48. Pekel, J.-F. et al. A near real-time water surface detection method based on HSV transformation of MODIS multi-spectral time series data. *Remote Sens. Environ.* **140**, 704–716 (2014).
49. Roberts, J. C. in *Coordinated and Multiple Views in Exploratory Visualization* (CMV'07 Fifth Int. Conf.) 61–71, <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4269947&isnumber=4269933> (IEEE, 2007).
50. Delaunay, B. Sur la sphere vide. *Bull. Acad. Sci. USSR* **7**, 793–800, <http://www.mathnet.ru/links/bf140e013bb2829a727614ee4e41051a/im4937.pdf> (1934).
51. Arendt, A. et al. *Randolph Glacier Inventory—A Dataset of Global Glacier Outlines: Version 5.0* <http://www.glims.org/RGI/> (Global Land Ice Measurements from Space, Digital Media, 2015).
52. Pesaresi, M. et al. A global human settlement layer from optical HR/VHR RS data: concept and first results. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **6**, 2102–2131 (2013).
53. Pesaresi, M. et al. *Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014* <http://publications.jrc.ec.europa.eu/repository/handle/JRC97705> (Publications Office of the European Union, 2016).
54. *Global 30-Arc Second Elevation Data Set (GTOPO30)* <https://lta.cr.usgs.gov/GTOPO30> (Department of the Interior, USGS, 1996).
55. Danielson, J. J. & Gesch, D. B. *Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010)*. USGS Report 2011–1073, <https://pubs.er.usgs.gov/publication/ofr20111073> (USGS Publications Warehouse, 2011).
56. Jarvis, A., Reuter, H. I., Nelson, A. & Guevara, E. *Hole-filled SRTM for the Globe Version 4* <http://srtm.csi.cgiar.org> (CGIAR-CSI SRTM 90m Database, 2008).
57. *Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global* <https://lta.cr.usgs.gov/SRTM1Arc> (Land Processes Distributed Active Archive Center (LP DAAC), USGS/EROS, accessed November 2016).
58. Zhu, Z. & Woodcock, C. E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **118**, 83–94 (2012).
59. Seipp, K., Ochoa, X., Gutiérrez, F. & Verbert, K. A research agenda for managing uncertainty in visual analytics. *Gesellsch. Inform.* 1–10 (Human Factors in Information Visualization and Decision Support Systems (HFIDSS), Mensch und Computer Workshopband, 2016).
60. *Shuttle Radar Topography Mission Water Body Data* https://lta.cr.usgs.gov/srtm_water_body_dataset (SRTM Water Body Data (SWBD), 2003).
61. *Global Administrative Areas (GADM) version 2.6*, <https://waterloo.ca/library/geospatial/collections/us-and-world-geospatial-data-resources/global-administrative-areas-gadm> (Univ. Berkeley, Museum of Vertebrate Zoology and the International Rice Research Institute, 2012).
62. Zhang, Y., Li, B. & Zheng, D. Datasets of the boundary and area of the Tibetan Plateau. *Glob. Change Res. Data Publ. Repository* <http://www.geodoi.ac.cn/weben/doi.aspx?id=135> (2014).
63. Papa, F. et al. Interannual variability of surface water extent at the global scale, 1993–2004. *J. Geophys. Res.* **115**, D12 (2010).
64. Klein, I. et al. Results of the Global WaterPack: a novel product to assess inland water body dynamics on a daily basis. *Remote Sens. Lett.* **6**, 78–87 (2015).
65. Belward, A. S. & Skøien, J. O. Who launched what, when and why; trends in Global Land-Cover Observation capacity from civilian Earth Observation satellites. *ISPRS J. Photogramm. Remote Sens.* **103**, 115–128 (2015).
66. Cohen, W. B. & Goward, S. N. Landsat's role in ecological applications of remote sensing. *Bioscience* **54**, 535–545 (2004).



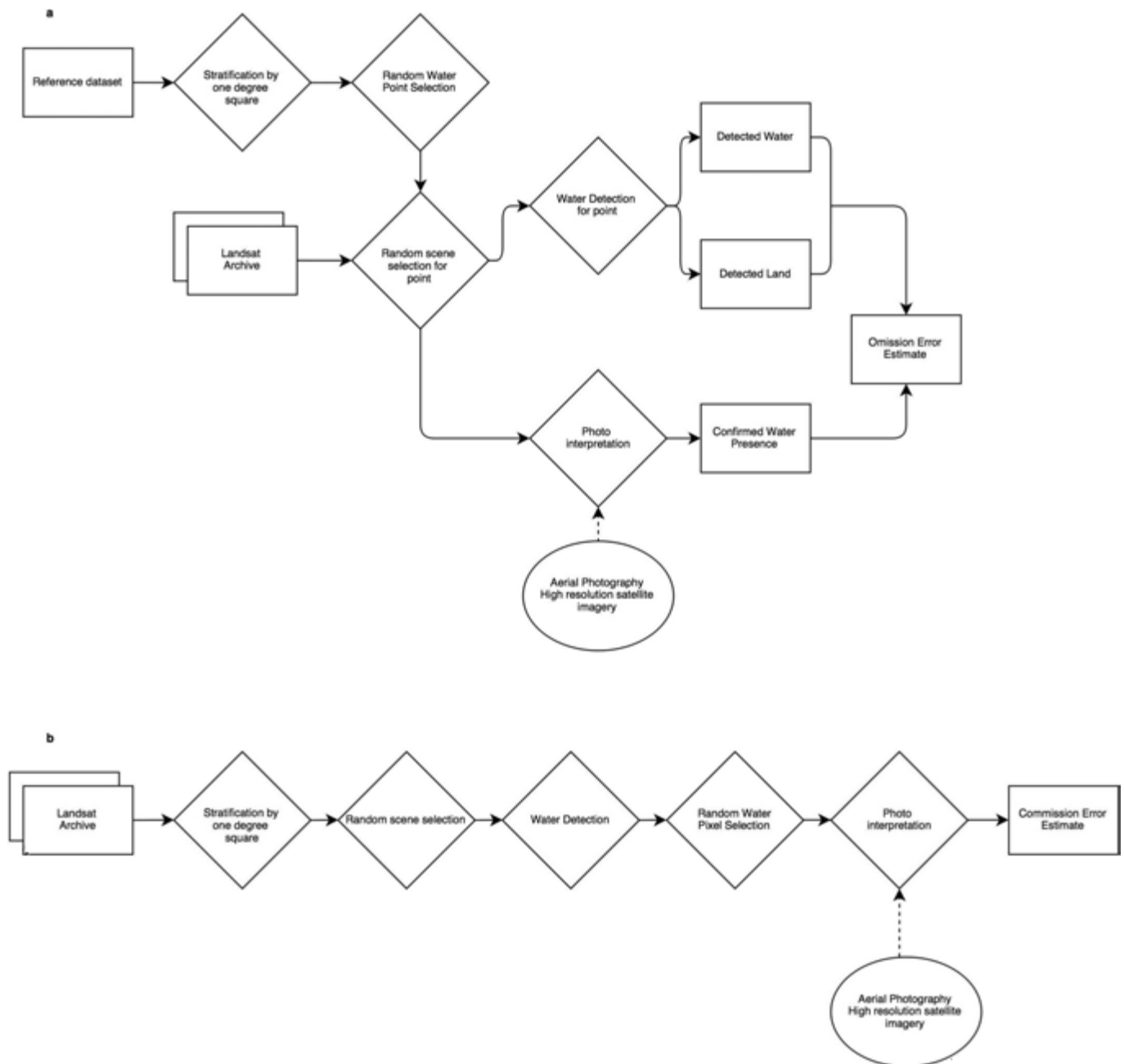
Extended Data Figure 1 | Geographic and temporal coverage of the Landsat 5, 7 and 8 L1T archive between 16 March 1984 and 10 October 2015.
a, Total number of unique views. **b**, First year of imaging. **c**, Number of scenes per month and year.



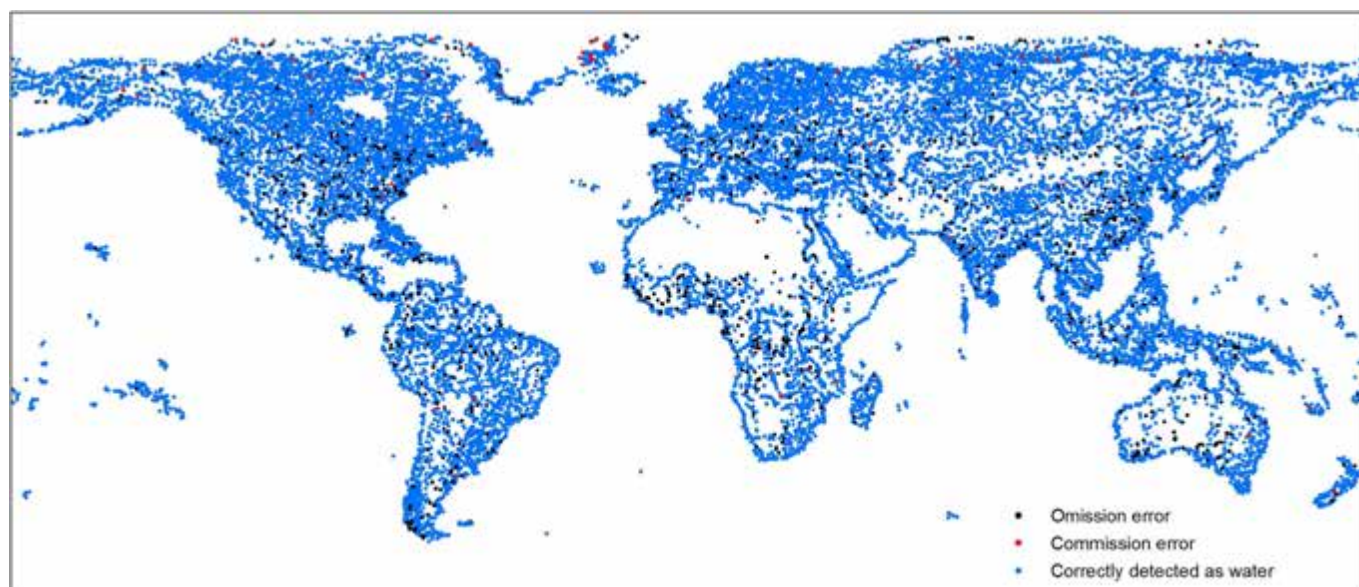
Extended Data Figure 2 | Diagram of the expert system classifier.



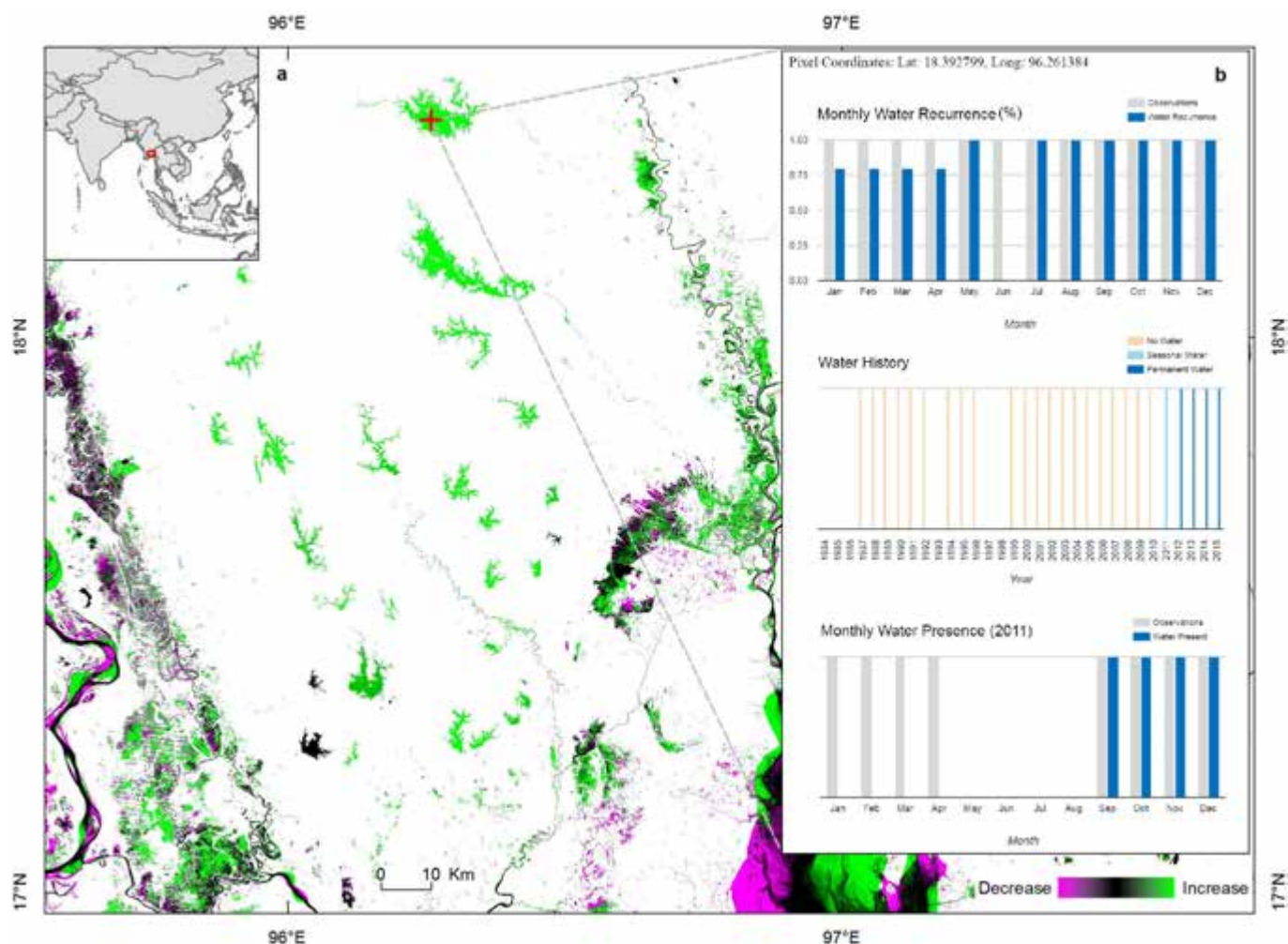
Extended Data Figure 3 | Multispectral feature-space occupied by water and other surfaces. **a**, Hue (from SWIR2, NIR, red) versus NDVI. **b**, Hue versus Value (both from SWIR2, NIR, red). **c**, Hue versus Saturation (both from SWIR2, NIR, red). **d**, Hue versus Value (both from NIR, green, blue).



Extended Data Figure 4 | Diagram of the validation protocol. a, Omission error protocol. b, Commission error protocol.

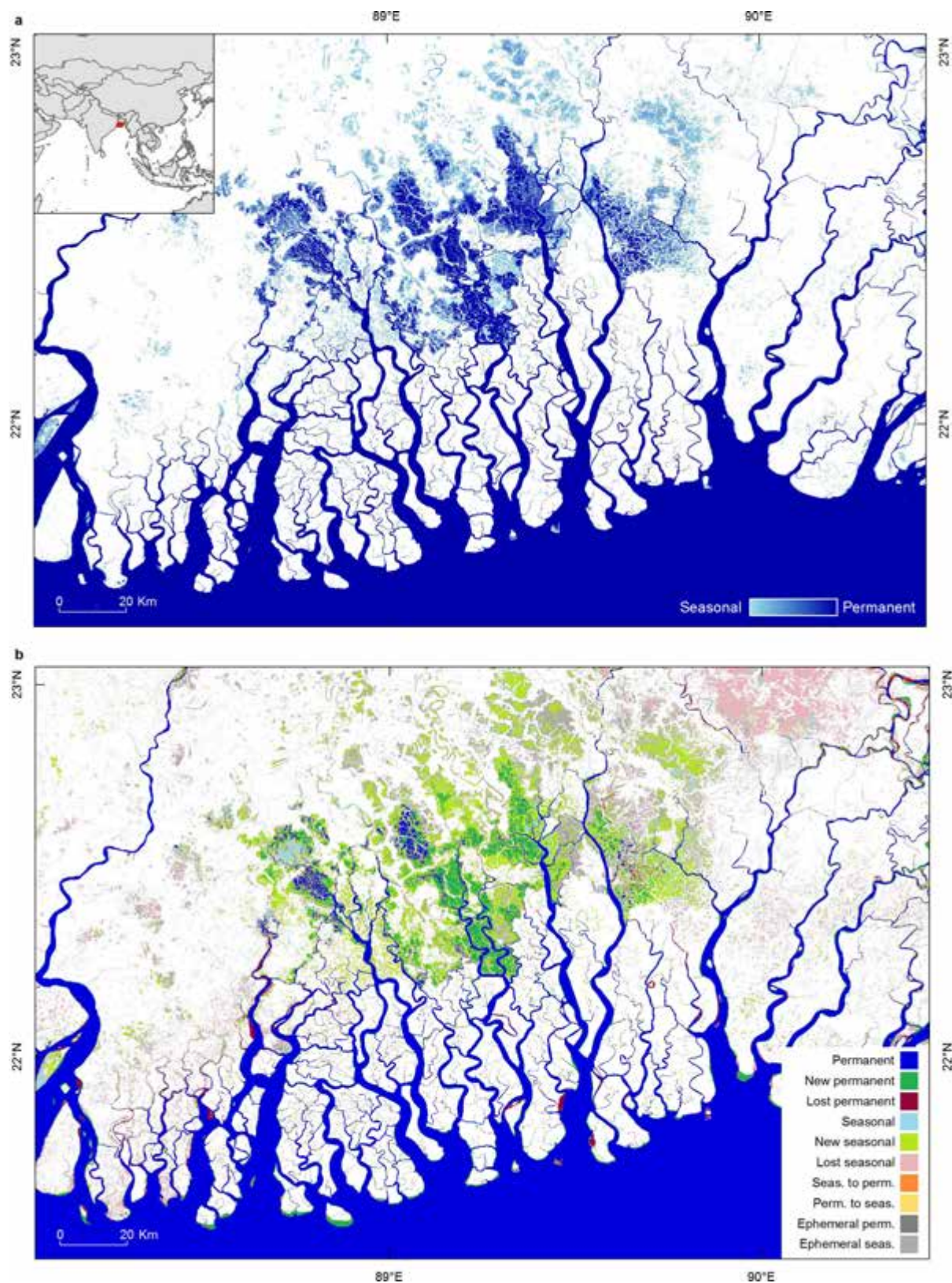


Extended Data Figure 5 | Global geographic distribution of validation sample points and error.



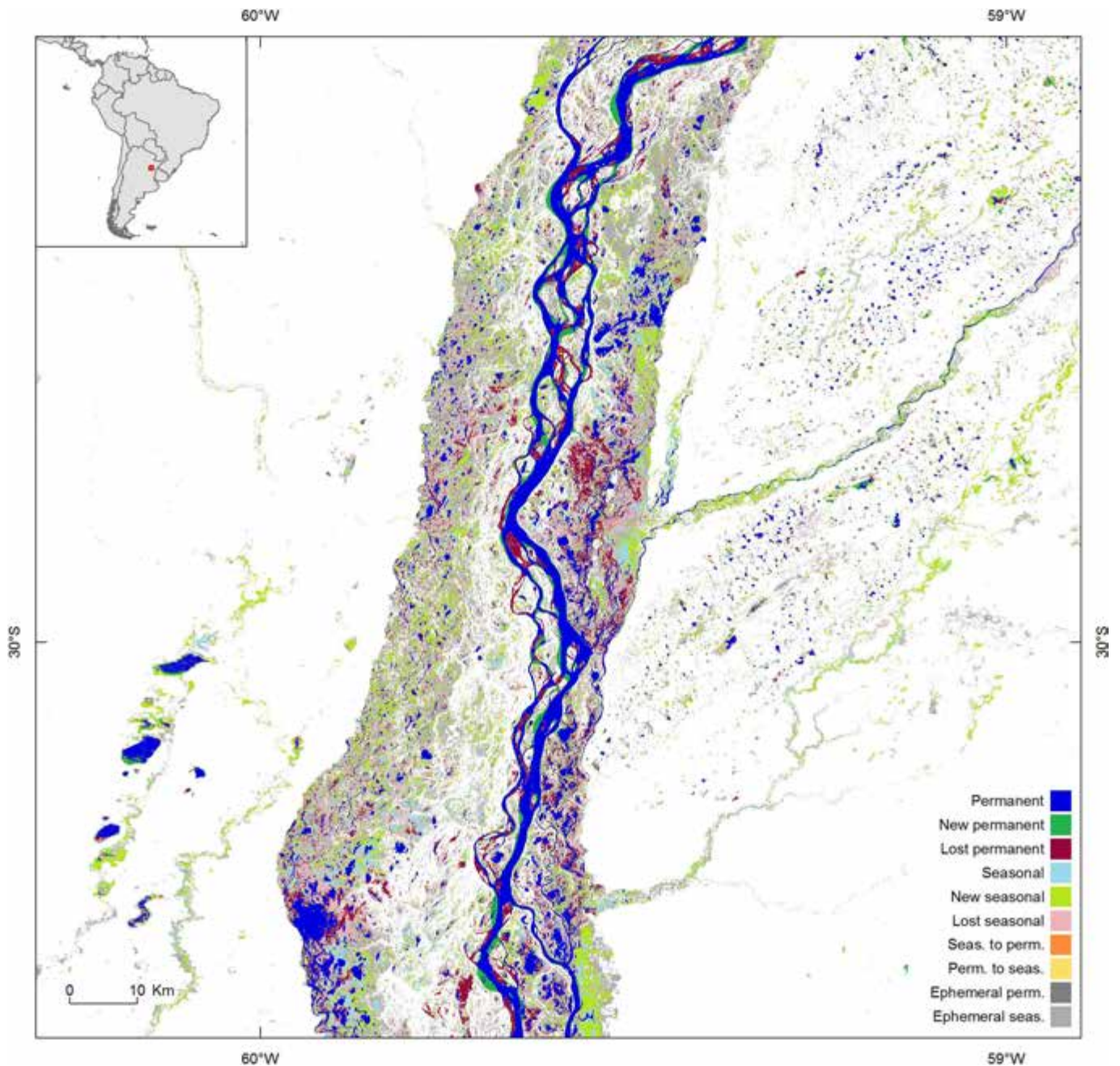
Extended Data Figure 6 | Mapping the history of surface water occurrence. **a**, Examples of increasing surface water occurrence in Myanmar (see inset for regional context). **b**, Pixel-based temporal profiles showing recurrence by month over 32 years (top), water history by seasonality class and by year over 32 years (middle) and monthly water presence for each year in the water seasonality record, in this case 2011 (bottom). Collectively, the graphs show that at this location (latitude

18.3928°, longitude 96.2633°) there are no valid observations available for the period 1984–1986, in 1993, 1997 or 1998 (the gaps in the middle graph), that before 2011 this was dry land, that the dam formed in 2011 and this point was flooded sometime between April and September (bottom), but since then it has been permanent water (centre), and that in the 32 years of observation water has not been detected in June (no observations have been made in June since the dam filled (top)).



Extended Data Figure 7 | Mapping changes in intra-annual persistence (seasonality). **a**, Surface water seasonality between October 2014 and October 2015 in the Sundarbans in Bangladesh (see inset). **b**, Changes in inter-annual persistence between 1984 and 2015. The increase in

permanent surface water at the expense of seasonal is indicative of changes in land use from seasonally flooded paddy fields to permanently flooded fishponds.



Extended Data Figure 8 | Water transitions map the Parana river. The regional context is shown in the insets. River channel migration, changes to seasonal water across the floodplain and transitions from permanent to seasonal water ('New seasonal') and seasonal to permanent water ('New permanent'), visible in the figure, are symptomatic of habitat fragmentation and changing ecosystem service delivery.

Extended Data Table 1 | Validation results

a)

	Landsat 5			Landsat 7			Landsat 8		
	Overall	Seasonal	Permanent	Overall	Seasonal	Permanent	Overall	Seasonal	Permanent
Misclassified as land #	18	6	12	20	7	13	30	10	20
Correctly classified as water #	3226	493	2733	3037	424	2613	6525	671	5854
Commission accuracy	99.45%	98.80%	99.56%	99.35%	98.38%	99.50%	99.54%	98.53%	99.66%

b)

	Landsat 5			Landsat 7			Landsat 8		
	Overall	Seasonal	Permanent	Overall	Seasonal	Permanent	Overall	Seasonal	Permanent
Misclassified as land #	233	146	87	343	172	171	425	336	89
Correctly classified as water #	7561	436	7125	7808	485	7323	10898	1151	9747
Omission accuracy	97.01%	74.91%	98.79%	95.79%	73.82%	97.72%	96.25%	77.40%	99.10%

Accuracy results judged against: **a**, commission error by sensor and by seasonality class; **b**, omission error by sensor and by seasonality class.

Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism

Neelroop N. Parikshak^{1,2*}, Vivek Swarup^{1,2*}, T. Grant Belgard^{1,2*†}, Manuel Irimia^{3,4}, Gokul Ramaswami^{1,2}, Michael J. Gandal^{1,2}, Christopher Hart^{1,2}, Virpi Leppä¹, Luis de la Torre Ubieta^{1,2}, Jerry Huang^{1,2}, Jennifer K. Lowe¹, Benjamin J. Blencowe^{5,6}, Steve Horvath^{7,8} & Daniel H. Geschwind^{1,2,7}

Autism spectrum disorder (ASD) involves substantial genetic contributions. These contributions are profoundly heterogeneous but may converge on common pathways that are not yet well understood^{1–3}. Here, through post-mortem genome-wide transcriptome analysis of the largest cohort of samples analysed so far, to our knowledge^{4–7}, we interrogate the noncoding transcriptome, alternative splicing, and upstream molecular regulators to broaden our understanding of molecular convergence in ASD. Our analysis reveals ASD-associated dysregulation of primate-specific long noncoding RNAs (lncRNAs), downregulation of the alternative splicing of activity-dependent neuron-specific exons, and attenuation of normal differences in gene expression between the frontal and temporal lobes. Our data suggest that SOX5, a transcription factor involved in neuron fate specification, contributes to this reduction in regional differences. We further demonstrate that a genetically defined subtype of ASD, chromosome 15q11.2–13.1 duplication syndrome (dup15q), shares the core transcriptomic signature observed in idiopathic ASD. Co-expression network analysis reveals that individuals with ASD show age-related changes in the trajectory of microglial and synaptic function over the first two decades, and suggests that genetic risk for ASD may influence changes in regional cortical gene expression. Our findings illustrate how diverse genetic perturbations can lead to phenotypic convergence at multiple biological levels in a complex neuropsychiatric disorder.

We performed rRNA-depleted RNA sequencing (RNA-seq) of 251 post-mortem samples of frontal and temporal cortex and cerebellum from 48 individuals with ASD and 49 control subjects (Methods and Extended Data Fig. 1a–h). We first validated differential gene expression (DGE) between samples of cortex from control individuals and those with ASD (ASD cortex) by comparing gene expression with that of different individuals from those previously profiled by microarray⁸, and found strong concordance ($R^2 = 0.60$; Fig. 1a, Extended Data Fig. 1i). This constitutes an independent technical and biological replication of shared molecular alterations in ASD cortex.

We next combined covariate-matched samples from individuals with idiopathic ASD to evaluate changes across the entire transcriptome. Compared to control cortex, 584 genes showed increased expression and 558 showed decreased expression in ASD cortex (Fig. 1b; Benjamini–Hochberg FDR < 0.05, linear mixed effects model; see Methods). This DGE signal was consistent across methods, unrelated to major confounders, and found in more than two-thirds of ASD samples (Extended Data Fig. 1j–m). We performed a classification analysis to confirm that gene expression in ASD could separate samples by disease

status (Extended Data Fig. 2a) and confirmed the technical quality of our data with qRT-PCR (Extended Data Fig. 2b, c). We next evaluated enrichment of the gene sets for pathways and cell types (Extended Data Fig. 2d, e), and found that the downregulated set was enriched in genes expressed in neurons and involved in neuronal pathways, including *PVALB* and *SYT2*, which are highly expressed in interneurons; by contrast, the upregulated gene set was enriched in genes expressed in microglia and astrocytes⁸.

Although there was no significant DGE in the cerebellum (FDR < 0.05, P distributions in Fig. 1b), similar to observations in a smaller cohort⁸, there was a replication signal in the cerebellum and overall concordance between ASD-related fold changes in the cortex and cerebellum (Extended Data Fig. 2f–h). The lack of significant DGE in the cerebellum is explained by the fact that changes in expression were consistently stronger in the cortex than in the cerebellum (Extended Data Fig. 2h), which suggests that the cortex is more selectively vulnerable to these transcriptomic alterations. We also compared our results to an RNA-seq study of protein coding genes in the occipital cortex of individuals with ASD and control subjects⁴. Despite significant technical differences that reduce power to detect DGE, and profiling of different brain regions in that study, there was a weak but significant correlation in fold changes, which was due mostly to upregulated genes in both studies ($P = 0.038$, Extended Data Fig. 2i, j).

We next explored lncRNAs, most of which have little functional annotation, and identified 60 lncRNAs in the DGE set (FDR < 0.05, Extended Data Fig. 2k). Multiple lines of evidence, including developmental regulation in RNA-seq datasets and epigenetic annotations, support the functionality of most of these lncRNAs (Supplementary Table 2). Moreover, 20 of these lncRNAs have been shown to interact with microRNA (miRNA)–protein complexes, and 9 with the fragile X mental retardation protein (FMRP), whose mRNA targets are enriched in ASD risk genes^{9,10}. As a group, these lncRNAs are enriched in the brain relative to other tissues (Extended Data Fig. 2l, m) and most that have been evaluated across species exhibit primate-specific expression patterns in the brain¹¹, which we confirm for several transcripts (Supplementary Information, Extended Data Fig. 3a–h). We highlight two primate-specific lncRNAs, *LINC00693* and *LINC00689*. Both interact with miRNA processing complexes and are typically downregulated during development¹², but are upregulated in ASD cortex (Fig. 1c, d, Extended Data Fig. 2n). These data show that dysregulation of lncRNAs, many of which are brain-enriched, primate-specific, and predicted to affect protein expression through miRNA or FMRP interactions, is an integral component of the transcriptomic signature of ASD.

¹Center for Autism Research and Treatment and Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California 90095, USA. ²Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, 695 Charles E. Young Drive South, Los Angeles, California 90095, USA.

³Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 88 Dr. Aiguader, Barcelona 08003, Spain. ⁴Universitat Pompeu Fabra (UPF), Barcelona, Spain. ⁵Donnelly Centre, University of Toronto, 160 College Street, Toronto, ON M5S 3E1, Canada. ⁶Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada.

⁷Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA. ⁸Department of Biostatistics, David Geffen School of Medicine, University of California, Los Angeles, California, USA. [†]Present address: Verge Genomics, 42A Dore Street, San Francisco, California 94103, USA.

*These authors contributed equally to this work.

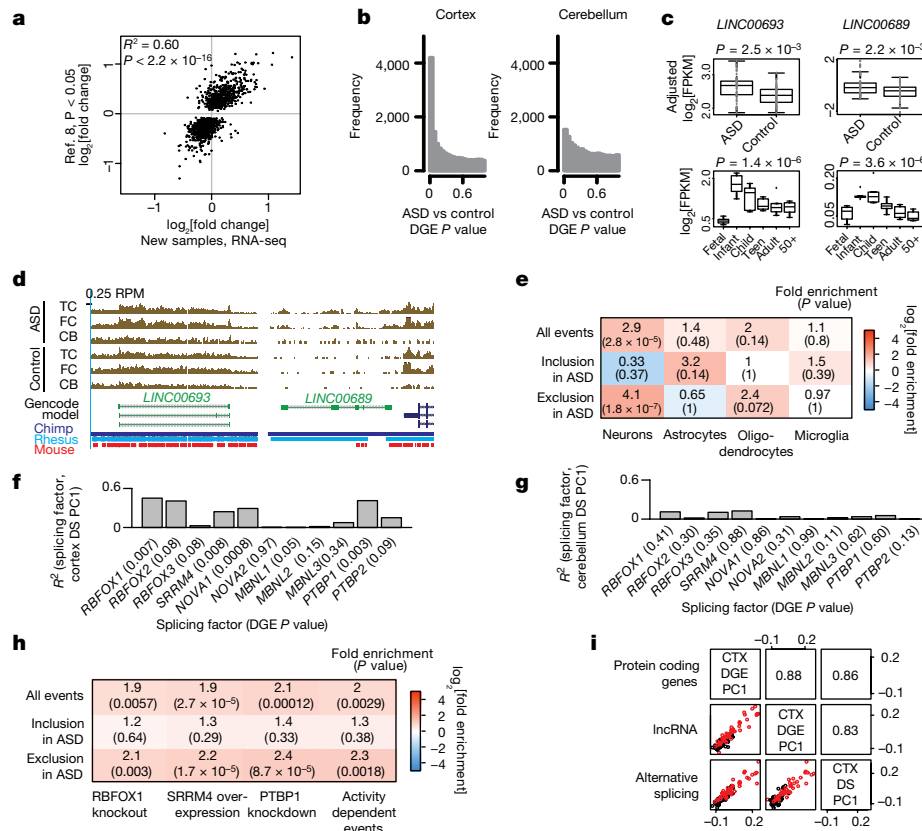


Figure 1 | Transcriptome-wide differential gene expression and alternative splicing in ASD. **a**, Replication of DGE between ASD and control cortex from previously analysed samples (16 ASD and 16 control on microarray⁸) with new age- and sex-matched cortex samples (15 ASD and 17 control). **b**, P value distribution of the linear mixed effect (LME) model DGE results for cortex and cerebellum. **c**, *LINC00693* and *LINC00689* are upregulated in ASD and downregulated during cortical development (developmental expression data from ref. 12). Two-sided ASD–control P values are computed by the LME model, developmental P values are computed by analysis of variance (ANOVA). FPKM, fragments per kilobase million mapped reads. **d**, UCSC genome browser track displaying reads per million (RPM) in ASD and control samples along with sequence

Previous studies have evaluated alternative splicing in ASD and its relation to specific splicing regulators in small sets of selected samples across individuals^{8,13,14}. Given the increased sequencing depth, reduced 5′–3′ sequencing bias, and larger cohort represented here, we were able to perform a comprehensive analysis of differential alternative splicing (Extended Data Fig. 4a). We found a significant differential splicing signal over background in the cortex (1,127 differential splicing events in 833 genes; Methods), but not in the cerebellum (P distributions in Extended Data Fig. 4b, c). We confirmed that confounders do not account for the differential splicing signal, reproduced the global differential splicing signal with an alternative pipeline¹⁵, and performed technical validation with RT–PCR (Extended Data Figs 4d–g, 5a), confirming the differential splicing analysis. Notably, the differential splicing molecular signature is not driven by DGE (Extended Data Fig. 4h), consistent with the observation that splicing alterations are related to common disease risk independently of gene expression changes¹⁶.

Cell-type specific enrichment and pathway analysis of alternative splicing demonstrated that most differential splicing events involve exclusion of neuron-specific exons¹⁷ (Fig. 1e, Extended Data Fig. 4i). Therefore, we next investigated whether the shared splicing signature in ASD could be explained by perturbations in splicing factors known to be important in nervous system function^{8,14} (Extended Data Fig. 4j), and found high correlations between splicing factor

conservation for *LINC00693* and *LINC00689*. **e**, Cell-type enrichment analysis of differential alternative splicing events from cortex using exons with ΔPSI (per cent spliced in) $>50\%$ in each cell type compared to the others¹⁷. **f**, **g**, Correlation between the first principal component (PC1) of the cortex differential splicing (DS) set and gene expression of neuronal splicing factors in cortex (**f**) and cerebellum (**g**) (DGE P value in parentheses). **h**, Enrichment among ASD differential splicing events and events regulated by splicing factors and neuronal activity (see Methods). **i**, Correlations between the PC1 across the ASD versus control analyses for different transcriptome subcategories. Bottom left: scatterplots of the principal components for ASD (red) and control (black) individuals. Top right: pairwise correlation values between principal components.

expression and differential splicing in the cortex (Fig. 1f) but not the cerebellum (Fig. 1g). The absence of neuronal splicing factor DGE or correlation with splicing changes in the cerebellum is consistent with the absence of a differential splicing signal in the cerebellum and suggests that these splicing factors contribute to cortex-biased differential splicing. Previous experimental perturbation of three splicing factors, *Rbfox1* (ref. 18), *SRRM4* (ref. 19), and *PTBP1* (ref. 20), shows strong overlap with the differential splicing changes found in ASD cortex, further supporting these predicted relationships (Fig. 1h, Extended Data Fig. 5b). Given that differential splicing events in ASD cortex overlap significantly with those that are targets of neuronal splicing factors, we hypothesized that some of these events may be involved in activity-dependent gene regulation. Indeed, differential splicing events were significantly enriched in those previously shown to be regulated by neuronal activity²¹ (Fig. 1h). This overlap supports a model of ASD pathophysiology based on changes in the balance of excitation and inhibition and in neuronal activity²² and suggests that alterations in transcript structure are likely to be an important component.

When we compared the first principal component across samples for protein coding DGE, lncRNA DGE and differential splicing, we found remarkably high correlations ($R^2 > 0.8$), indicating that molecular convergence is likely to be a unitary phenomenon across multiple levels of transcriptome regulation in ASD (Fig. 1i).

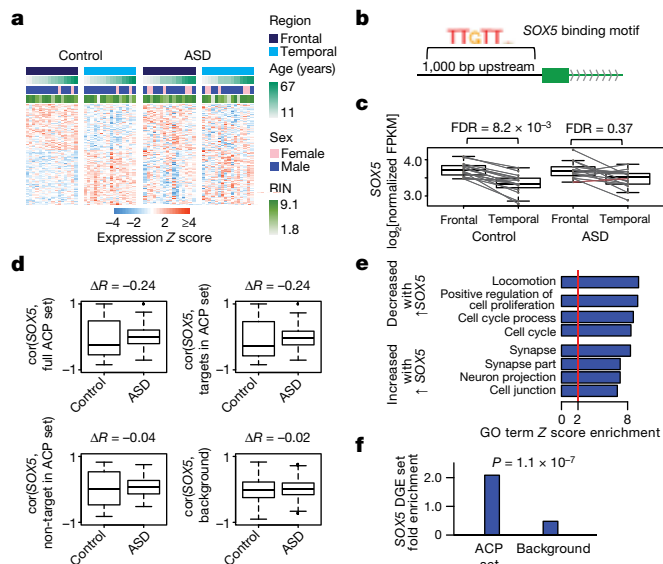


Figure 2 | Attenuation of cortical patterning in ASD. **a**, Heat map of genes exhibiting DGE between frontal and temporal cortex at $FDR < 0.05$. In control cortex and ASD cortex, 551 genes and 51 genes, respectively, show DGE in frontal versus temporal cortex. The ACP set is defined as the 523 genes that show DGE between regions in control but not ASD samples. RIN, RNA integrity number. **b**, Schematic of transcription factor motif enrichment upstream of genes in the ACP set. **c**, SOX5 exhibits attenuated cortical patterning in ASD (lines: frontal–temporal pairs from the same individual). **d**, Correlation between SOX5 expression and predicted targets in control and ASD samples for all ACP genes (top left), SOX5 targets from the ACP set (top right), SOX5 non-targets from the ACP set (bottom left), and background (all other genes, bottom right). Plots show the distribution of Pearson correlation values between SOX5 and other genes in ASD and control samples. ΔR , change in median R value between distributions. **e**, Gene Ontology (GO) term enrichment for genes upregulated and downregulated after SOX5 overexpression in neural progenitor cells. **f**, Enrichment analysis of the SOX5 differential gene expression (DGE) set in the ACP set and all other genes (background). P represents significance in enrichment over background by two-sided Fisher's exact test.

Previous analysis suggested that the typical pattern of transcriptional differences between the frontal and temporal cortices may be attenuated in ASD⁸. We confirmed this in our larger cohort and identified 523 genes that differed significantly in expression between the frontal

cortex and the temporal cortex in control subjects, but not those with ASD (Fig. 2a); we refer to these genes as the ‘attenuated cortical patterning’ (ACP) set (Extended Data Fig. 6a). We demonstrated the robustness of attenuation in cortical patterning in ASD by confirming that the ACP set was not more variable than other genes, that attenuation of cortical patterning was robust to removal of previously analysed samples⁸, and that the effect could also be observed using a different classification approach (Extended Data Fig. 6b–h).

Pathway and cell-type analysis showed that the ACP set is enriched in *Wnt* signalling, calcium binding, and neuronal genes (Extended Data Fig. 6i, j, Supplementary Information). We next explored potential regulators of cortical patterning by transcription factor binding site enrichment (Extended Data Fig. 6k). Among the transcription factors identified, SOX5 was of particular interest because of its known role in mammalian corticogenesis^{23,24}, its sole membership in the ACP set, and its correlation with predicted targets in the brains of control subjects, which is lost in ASD (Fig. 2b–d). We confirmed that a significant proportion of ACP genes are regulated by SOX5 by overexpressing it in human neural progenitors. SOX5 induced synaptic genes and repressed cell proliferation (Fig. 2e), and predicted SOX5 targets exhibited net down-regulation, consistent with the repressive function of SOX5 (Fig. 2f, Extended Data Fig. 6l, m). These findings support the prediction that attenuated patterning of the transcription factor SOX5 between cortical regions contributes to direct alterations in patterning of SOX5 targets.

We also evaluated DGE and differential splicing in nine individuals with dup15q (which is among the most common and penetrant forms of ASD) and independent controls (Extended Data Fig. 7a, b). Significant upregulation in the 15q11.1–13.2 region (*cis*) was evident in duplication carriers, but not in idiopathic ASD (Fig. 3a). Remarkably, genome-wide (*trans*) DGE and differential splicing patterns were highly concordant between dup15q and ASD (Fig. 3b, c, Extended Data Fig. 7c–e). Moreover, alterations in dup15q cortex were of greater magnitude and more homogeneous than those observed in idiopathic ASD cortex (Fig. 3d, Extended Data Fig. 7f, g). Analysis of DGE in the cerebellum confirmed a weaker signal than in the cortex and demonstrated that *cis* changes in dup15q cerebellum (Extended Data Fig. 7h–j) were more concordant with the cortex than *trans* changes (Extended Data Fig. 7k, l), further supporting the observation that the cortex is selectively vulnerable to transcriptomic alteration in ASD. Together, the DGE and differential splicing analyses in dup15q provide further biological validation of the ASD transcriptomic signature and demonstrate that a genetically defined form of ASD exhibits similar changes to idiopathic ASD.

We next applied weighted gene co-expression network analysis (WGCNA; Methods) and evaluated the biological functions and ASD

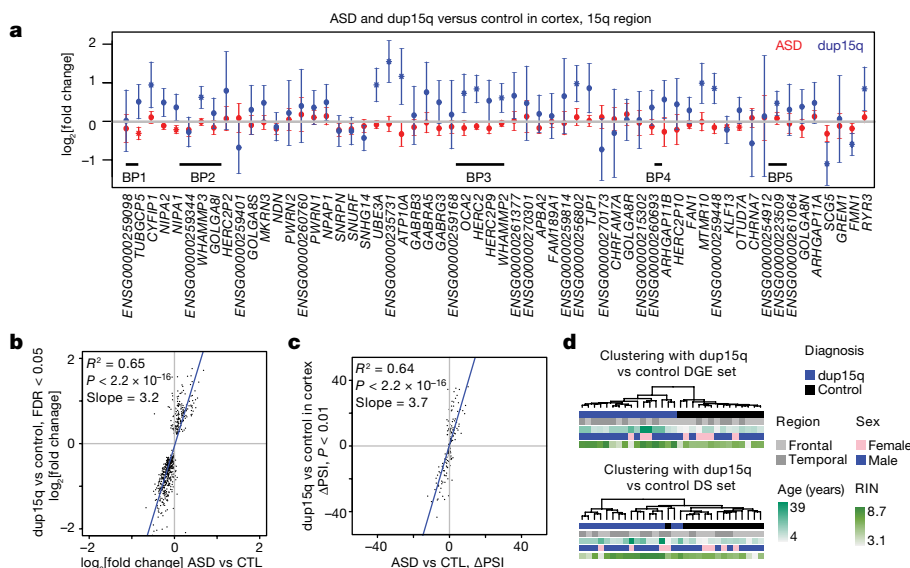


Figure 3 | Duplication 15q syndrome recapitulates transcriptomic changes in idiopathic ASD. **a**, DGE changes across the 15q11–13.2 region for ASD and dup15q compared to control. Error bars show 95% confidence intervals for the fold changes. * $FDR < 0.05$ across this region. BP, breakpoint. **b**, Comparison of DGE effect sizes in dup15q versus control and ASD versus control. **c**, Comparison of differential alternative splicing effect sizes in dup15q versus control and ASD versus control. **d**, Average linkage hierarchical clustering of dup15q samples and controls using the DGE and differential alternative splicing (DS) gene sets.

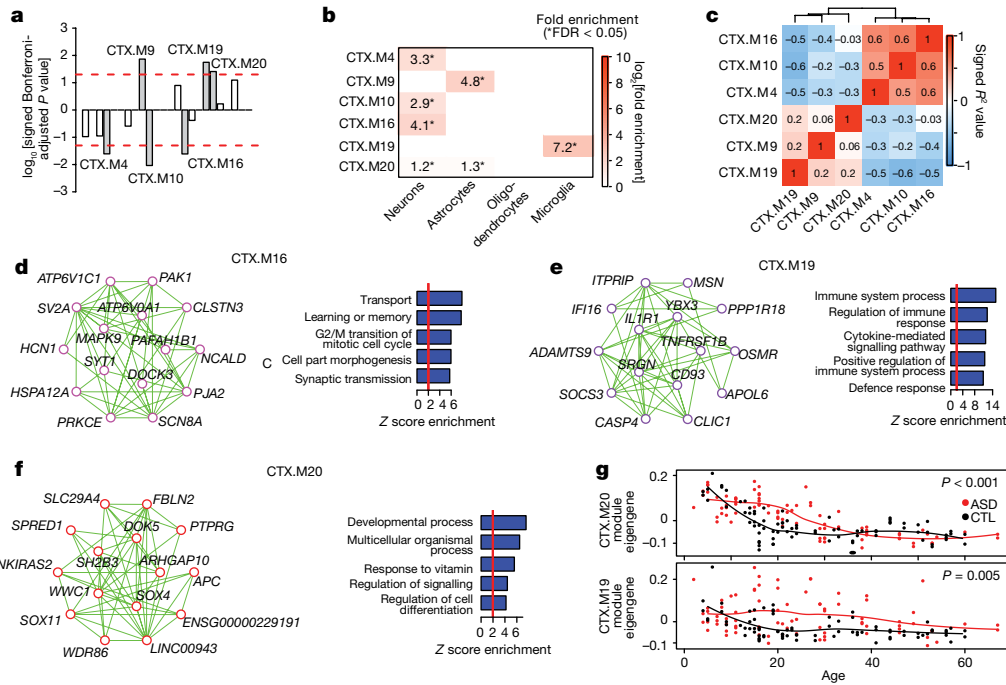


Figure 4 | Co-expression network analysis. **a**, Signed association of module eigengenes with diagnosis (Bonferroni-corrected P value from an LME model, see Extended Data Fig. 8c and Methods). Positive values indicate modules with an increased expression in ASD samples. Grey bars with labels signify six ASD-associated modules. **b**, Cell-type enrichment for the ASD-associated modules. **c**, Heat map of correlations between

association of the 24 co-expression modules identified (Extended Data Fig. 8a–d). Of the six modules associated with ASD, three were upregulated and three were downregulated, and each showed significant cell-type enrichment (Fig. 4a, b). This analysis corroborates and extends previous work by identifying sub-modules of those previously identified, thus demonstrating greater biological specificity (Extended Data Figs 8e, 9a). It also confirms that downregulated modules are enriched in synaptic function and neuronal genes, that upregulated modules are enriched in genes associated with inflammatory pathways and glial function^{4,8}, and that microglial and synaptic modules exhibit significant anticorrelation (Fig. 4c). Furthermore, the downregulated modules CTX.M10 and CTX.M16 are enriched in genes previously related to neuronal firing rate, consistent with the overlap of dysregulated splicing with events regulated by neuronal activity (Extended Data Fig. 9b and Fig. 1h). One glial and one neuronal module are highlighted in Fig. 4d, e (the remainder in Extended Data Fig. 9c–e). Remarkably, the upregulated module CTX.M20 was not found in previous analyses, overlaps significantly with the ACP set (FDR < 0.05, Extended Data Fig. 9a), and contains genes implicated in development and regulation of cell differentiation (Fig. 4f).

We also leveraged our large sample and younger age-matched ASD and control samples to detect differences in developmental trajectories in ASD compared to control subjects. We identified a remarkable difference in CTX.M19 and CTX.M20 during the first two decades of life (Fig. 4g, additional age trajectories in Extended Data Fig. 9f) that is most consistent with an evolving process during early brain development that stabilizes starting in late childhood and early adolescence. We also found preservation of most cortex modules in the cerebellum, but with weaker associations to ASD (Extended Data Fig. 10a–h, Supplementary Table 4), consistent with the DGE analysis showing that ASD-related changes are substantially smaller in the cerebellum.

To determine the role of genetic factors in transcriptomic dysregulation, we evaluated enrichment in genes affected by ASD-associated rare mutations and common variants (Extended Data Fig. 9a). One module,

ASD-associated module eigengenes sorted by average linkage hierarchical clustering. **d–f**, Module plots displaying the top 15 hub genes and top 50 connections along with the GO term enrichment of each module. **g**, Plot of CTX.M20 and CTX.M19 module eigengenes across age. P values are for the difference between temporal trajectories for ASD and control by permutation test (see Methods).

CTX.M24, exhibited significant enrichment for rare mutations found in ASD, while rare *de novo* mutations associated with intellectual disability were most strongly enriched in CTX.M22 (FDR < 0.05, Extended Data Fig. 9a). Remarkably, CTX.M24 was significantly enriched for lncRNAs, genes expressed highly during fetal cortical development, and genes harbouring protein-disrupting mutations found in ASD, suggesting that lncRNAs will be important targets for investigation in ASD^{10,25} (FDR < 0.05, Extended Data Fig. 9a, g). By contrast, enrichment for ASD-associated common variation was observed in CTX.M20 (FDR < 0.1, Extended Data Fig. 9h–i, Methods). As CTX.M20 is enriched for the ACP gene set, this suggests a potential link between polygenic risk and regional attenuation of gene expression in ASD. Several other ASD-associated modules showed a weaker common variant signal for ASD, including CTX.M16, which also shows a signal for schizophrenia polygenic risk. However, other phenotypes with larger, better-powered genome-wide association studies (GWAS) also demonstrate enrichment (Extended Data Fig. 9h–i). It will be necessary to perform this analysis with larger ASD GWAS in the future to fully understand the extent and specificity of the contribution of common variation to the transcriptome alterations in ASD.

These data contribute to a consistent emerging picture of the molecular pathology of ASD^{4,7,8,10,25–27}. Parsimony suggests that the highly overlapping expression pattern shared by individuals with dup15q and the majority of those with idiopathic ASD represents an evolving adaptive or maladaptive response to a primary insult rather than a secondary environmental hit. Although we observe no significant association of the ASD-associated transcriptome signature with either clinical or technical confounders, some of the changes are likely to represent consequences or compensatory responses, rather than causal factors. In this regard, it is notable that the observed transcriptome changes are consistent with an ongoing process that is triggered largely by genetic and prenatal factors^{3,9,10,23}, but that evolves during the first decade of brain development.

We interpret these data to suggest that aberrant microglia–neuron interactions reflect an early alteration in developmental trajectory

that becomes more evident in late childhood. This corresponds to the period of synapse elimination and stabilization after birth in humans^{28,29}, which may have significant implications for intervention. Our analyses also reveal primate-specific lncRNAs that are probably relevant to understanding human higher cognition^{11,30}. Co-expression of lncRNAs with genes harbouring ASD-associated protein coding mutations suggests that these noncoding RNAs are involved in similar biological functions and are potential candidate ASD risk loci. As future investigations pursue the full range of causal genetic variation that contributes to ASD risk, these data will be valuable for interpreting genetic and epigenetic studies of ASD and the relationship between ASD and other neuropsychiatric disorders.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 May; accepted 7 November 2016.

Published online 5 December 2016.

- Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
- Gratten, J., Visscher, P. M., Mowry, B. J. & Wray, N. R. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat. Genet.* **45**, 234–238 (2013).
- de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22**, 345–361 (2016).
- Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5**, 5748 (2014).
- Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiol. Dis.* **30**, 303–311 (2008).
- Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E. & Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* **57**, 1618–1628 (2001).
- Chow, M. L. *et al.* Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes at young versus mature ages. *PLoS Genet.* **8**, e1002592 (2012).
- Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
- Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
- Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
- Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
- Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat. Neurosci.* **18**, 154–161 (2015).
- Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Reports* **6**, 1139–1152 (2014).
- Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511–1523 (2014).
- Wu, J., Ancukow, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. Olego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.* **41**, 5149–5163 (2013).
- Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
- Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* **34**, 11929–11947 (2014).
- Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–1442 (2013).
- Raj, B. *et al.* A global regulatory mechanism for activating an exon network required for neurogenesis. *Mol. Cell* **56**, 90–103 (2014).
- Gueroussou, S. *et al.* An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* **349**, 868–873 (2015).
- Maze, I. *et al.* Critical role of histone turnover in neuronal transcription and plasticity. *Neuron* **87**, 77–94 (2015).
- Mullins, C., Fishell, G. & Tsien, R. W. Unifying views of autism spectrum disorders: a consideration of autoregulatory feedback loops. *Neuron* **89**, 1131–1156 (2016).
- Kwan, K. Y. *et al.* SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proc. Natl Acad. Sci. USA* **105**, 16021–16026 (2008).
- Lamb, A. N. *et al.* Haploinsufficiency of SOX5 at 12p12.1 is associated with developmental delays with prominent language delay, behavior problems, and mild dysmorphic features. *Hum. Mutat.* **33**, 728–740 (2012).
- Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).
- Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
- Blumenthal, I. *et al.* Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. *Am. J. Hum. Genet.* **94**, 870–883 (2014).
- Huttenlocher, P. R. Morphometric study of human cerebral cortex development. *Neuropsychologia* **28**, 517–527 (1990).
- Khundrakpam, B. S., Lewis, J. D., Zhao, L., Chouinard-Decorte, F. & Evans, A. C. Brain connectivity in normally developing children and adolescents. *Neuroimage* **134**, 192–203 (2016).
- Zhang, Y. E., Landback, P., Vbranovski, M. D. & Long, M. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* **9**, e1001179 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements Tissue, biological specimens or data used in this research were obtained from the Autism BrainNet (formerly the Autism Tissue Program), which is sponsored by the Simons Foundation, and the University of Maryland Brain and Tissue Bank, which is a component of the NIH NeuroBioBank. We are grateful to the patients and families who participate in the tissue donation programs. The authors acknowledge R. Zielke, J. Cottrell and R. Johnson, who assisted with sample acquisition from the latter brain bank. Funding for this work was provided by grants to D.H.G. (NIMH 5R37 MH060233, 5R01 MH09714 and 5R01 MH100027), N.N.P. (NRSA F30 MH099886, UCLA Medical Scientist Training Program), V.L. (Sigrid Juselius Fellowship) and T.G.B. (training grant 5T32 MH073526). Additional grants supporting this work include those to B.J.B. (CIHR, Alzheimer's Research Foundation and University of Toronto McLaughlin Centre) and M.I. (ERC-StG-LS2-637591). We also thank D. Polioudakis for assistance with data management and V. Chandran for discussion of transcription factor binding site analysis and providing software.

Author Contributions N.N.P. and D.H.G. planned and directed experiments, guided analyses, and wrote the manuscript with assistance from all authors. N.N.P., V.S. and T.G.B. performed dissections, RNA-seq analysis, and differential gene expression analysis. N.N.P. and V.S. performed splicing analysis. M.I. and B.J.B. provided splicing validation data and assisted with splicing analysis. N.N.P., V.S., S.H., G.R., M.J.G. and C.H. performed co-expression network analysis. N.N.P., T.G.B., V.L. and J.K.L. performed analysis of duplication 15q syndrome samples. V.S. performed RT-PCR validation experiments and V.S., L.d.I.T.U. and J.H. performed SOX5 validation experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.H.G. (dhg@mednet.ucla.edu).

Reviewer Information Nature thanks K. Mirnics and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Brain tissue. Human brain tissue for ASD and control individuals was acquired from the Autism Tissue Program (ATP) brain bank at the Harvard Brain and Tissue Bank (which has since been incorporated into the Autism BrainNet) and the University of Maryland Brain and Tissue Bank, a Brain and Tissue Repository of the NIH NeuroBioBank. Sample acquisition protocols were followed for each brain bank, and samples were de-identified before acquisition. Brain sample and donor metadata are available in Supplementary Table 1 and further information about samples can be found in the Supplementary Information. No statistical methods were used to predetermine sample size. The sample dissections, RNA extractions, and RNA sequencing experiments were randomized (Supplementary Information). The investigators were blinded to diagnosis until the analysis but unblinded during the analysis.

RNA library preparation, sequencing, mapping and quantification. A detailed protocol, including parameters given to programs for each step, is provided in the Supplementary Information. Briefly, starting with total RNA, rRNA was depleted (RiboZero Gold, Illumina) and libraries were prepared using the TruSeq v2 kit (Illumina) to construct unstranded libraries with a mean fragment size of 150 bp. Libraries underwent 50-bp paired end sequencing on an Illumina HiSeq 2000 or 2500 machine. Paired end reads were mapped to hg19 using Gencode v18 annotations³¹ via Tophat2 (ref. 32). Gene expression levels were quantified using union exon models with HTSeq³³. This approach counts only reads on exons or reads spanning exon–exon junctions, and is globally similar to including reads on the introns (whole gene model) or computing probabilistic estimates of expression levels (Extended Data Fig. 1e–g).

Differential gene expression. DGE analysis was performed with expression levels normalized for gene length, library size, and G+C content (referred to as 'normalized FPKM'). Cortex samples (frontal and temporal) were analysed separately from cerebellum samples. An LME model framework was used to assess differential expression in \log_2 [normalized FPKM] values for each gene for cortical regions because multiple brain regions were available from the same individuals. The individual donor identifier was treated as a random effect, and age, sex, brain region and diagnoses were treated as fixed effects. In the cerebellum DGE analysis, a linear model was used and brain region was not included as a covariate, because only one brain region was available in each individual and a handful of technical replicates could be removed for DGE analysis. We also used technical covariates accounting for RNA quality and batch effects as fixed effects in this model (Supplementary Information). Significant results are reported at Benjamini–Hochberg FDR < 0.05 (ref. 34), and full results are available in Supplementary Table 2.

Throughout the study, we assessed replication between datasets by evaluating the concordance between independent sample sets by comparing the squared correlation (R^2) of fold changes of genes in each sample set at a defined statistical cut-off. We set the statistical cut-off in one sample set (the y axis in the scatterplots) and computed the R^2 with fold changes in these genes in the comparator sample set (the x axis in the scatterplots). For details of the regularized regression analyses and cortical patterning analyses, see Supplementary Information.

Differential alternative splicing. Alternative splicing was quantified using the per cent spliced in (PSI) metric using Multivariate Analysis of Transcript Splicing (MATS, v3.08)³⁵. For each event, MATS reports counts supporting the inclusion (I) or splicing (S) of an event. To reduce spurious events due to low counts, we required at least 80% of samples to have $I + S \geq 10$. For these events, the PSI is calculated as $PSI = I/(I + S)$ (Extended Data Fig. 4a). Statistical analysis for differential alternative splicing was performed using the linear mixed effects model as described above for DGE; significant results are reported at Benjamini–Hochberg FDR < 0.5 (ref. 34). Full differential alternative splicing results are available in Supplementary Table 3.

Quantitative real-time PCR validation. In order to ensure that our RNA-seq data were high quality and our DGE models were accurate, we evaluated gene expression changes in a representative subset of four ASD and four control samples (Extended Data Fig. 2b). One microgram of total RNA was reverse-transcribed using Invitrogen Superscript IV reverse-transcriptase and oligo-dT primers (Invitrogen). Real-time PCR was performed on a Lightcycler 480 thermocycler in 10 μ l volume containing SYBR Green Master Mix (Roche) and gene-specific primers at a concentration of 0.5 mM each. The results shown in Extended Data Fig. 2c represent at least two independent cDNA synthesis experiments for each gene. *GAPDH* levels were used as an internal control.

For differential alternative splicing analysis, we validated selected events with semiquantitative RT–PCR using the same samples used for DGE validation. Total RNA (600 ng) was reverse-transcribed using Invitrogen Superscript IV reverse transcriptase and gene/exon-specific primers. cDNA (50 ng) was amplified by 25 cycles using PCR. PCR products were resolved on 3% high-resolution Metaphor agarose gels (Lonza) and counterstained with SYBR Gold for visualization

(Extended Data Fig. 5a, Supplementary Fig. 1). Gels were quantified using ImageJ (NIH).

Notably, this sample size is underpowered to evaluate significant changes in many genes or splicing events; however, the goal was to validate the accuracy of our data and analyses across genes, so we show the correlation of fold changes between ASD and control across genes or events. Genes and events were selected on the basis of being top hits or of particular biological interest. Sample details and primers are reported in Supplementary Tables 2 and 3.

Duplication 15q syndrome samples and analyses. For dup15q samples, the type of duplication and copy number in the breakpoint 2–3 region were available from previous work³⁶. To expand this to the regions between each of the recurrent breakpoint in these samples, eight out of nine dup15q brains were genotyped (one was not genotyped owing to limited tissue availability). The number of copies between each of the breakpoints is reported in Extended Data Fig. 7a. DGE and differential alternative splicing analysis for this set was performed with independent control samples from the main analysis, though the results were similar to those obtained using the larger set of controls used in the main analysis (Extended Data Fig. 7d, e).

Co-expression network analysis. The R package weighted gene co-expression network analysis (WGCNA) was used to construct co-expression networks using normalized data after adjustment to remove variability from technical covariates^{37,38} (Supplementary Information). We used the biweight midcorrelation to assess correlations between \log_2 [adjusted FPKM] and parameters for network analysis are described in Supplementary Information. Notably, we used a modified version of WGCNA that involves bootstrapping the underlying dataset 100 times and constructing 100 networks. The consensus of these networks (median edge strength across all bootstrapped networks) was then used as the final network³⁹, ensuring that a subset of samples does not drive the network structure.

For module-trait analyses, the first principal component of each module (the module eigengene³⁷) was related to ASD diagnosis, age, sex, and brain region with an LME model as above. These associations were also supported by enrichment analyses with ASD DGE genes in Extended Data Fig. 9a. Given that modules are relatively uncorrelated to each other, significant eigengene-trait results are reported at Bonferroni-corrected $P < 0.05$.

Module temporal trajectories were computed with the LOESS function in R. For both ASD and control samples, the function was used to create quartic splines on module eigengenes (degree = 2, span = 2/3). The trend difference statistic was taken as the largest difference between these fitted curves between the ages of 5 and 25 years. P values were computed using 5,000 permutations. Specifically, ASD and control labels were randomly permuted 5,000 times and splines were fit to the permuted groups; therefore, significant P values reject the null hypothesis of no relationship between age trends and disease status. Detailed statistics for module membership are available in Supplementary Table 2 and additional characterization of modules is available in Supplementary Table 4.

Enrichment analysis of gene sets and common variation. Gene set enrichment analyses were performed with a two-sided Fisher's exact test (cell type and splicing factor enrichments) or with logistic regression (Extended Data Fig. 9a, Supplementary Information). Results were corrected for multiple comparisons by the Benjamini–Hochberg method³⁴ when a large number of comparisons were performed.

GO term enrichment analysis was performed using GO Elite⁴⁰ with 10,000 permutations, and results are presented as enrichment Z scores. We present only the top molecular function and biological process terms for display purposes. Notably, for splicing analysis, we evaluated GO term enrichment by using the genes containing differential splicing alterations to identify functional enrichment. It is possible that longer genes, which contain more exons, also contain more detected splicing events. This could bias pathway and cell type enrichment to more neuronal and synaptic genes, which are, on average, longer than other genes in the genome. However, the correlation between the number of detected events in genes and gene length is minimal ($R^2 = 0.004$), and the correlation is even smaller for events at $P < 0.01$ ($R^2 = 0.00012$) demonstrating that longer genes are not more likely to contain differential splicing events.

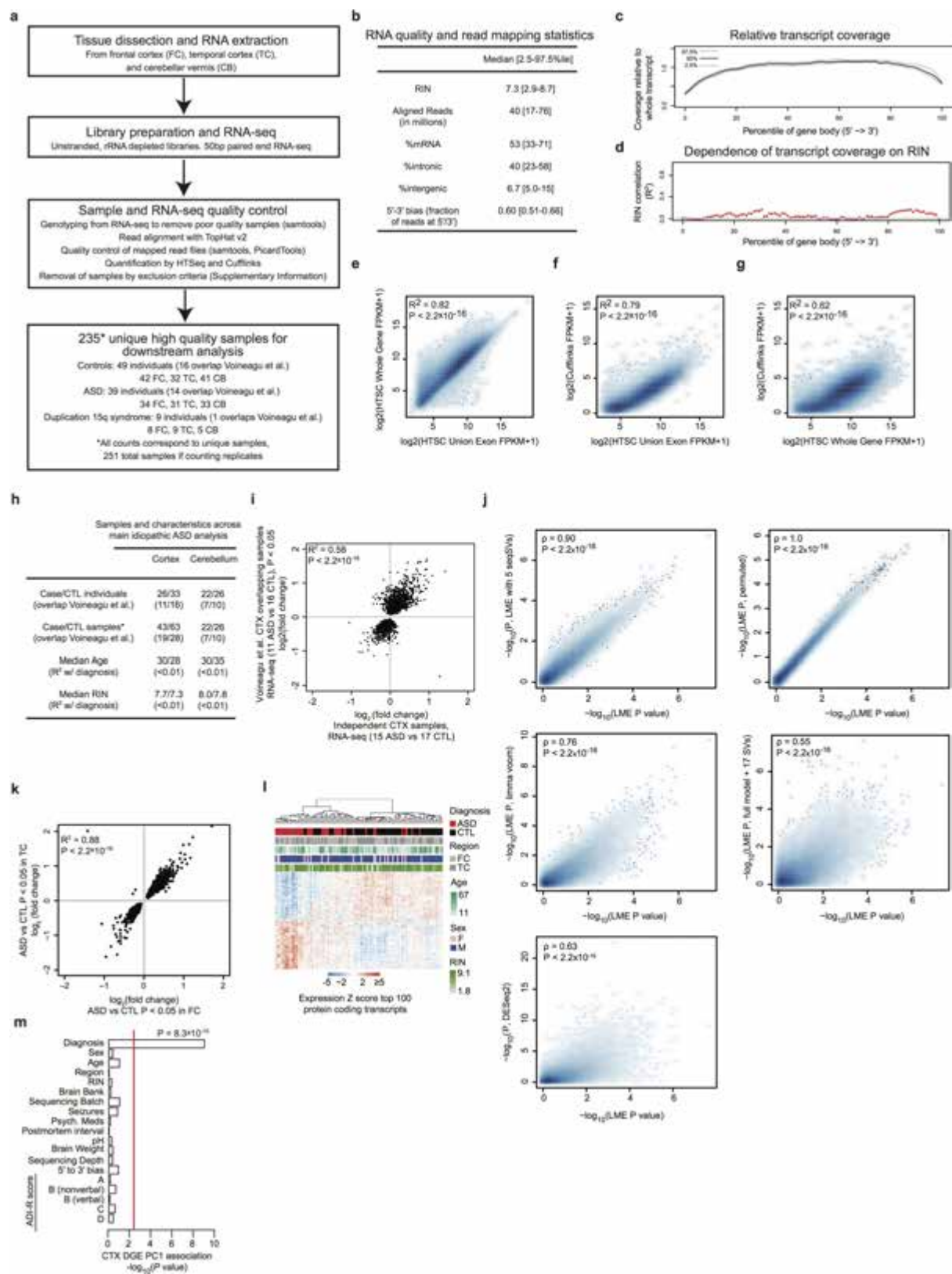
Common variant enrichment was evaluated by analysis of genome-wide association study (GWAS) signal with stratified linkage disequilibrium (LD) score regression to partition disease heritability within functional categories represented by gene co-expression modules⁴¹. This method uses GWAS summary statistics and LD explicitly modelled from an ancestry-matched 1,000 genomes reference panel to calculate the proportion of genome-wide single nucleotide polymorphism (SNP)-based heritability that can be attributed to SNPs within explicitly defined functional categories. To improve accuracy, these categories were added to a 'full baseline model' that includes 53 functional categories capturing a broad set of genomic annotations, as previously described⁴². Enrichment is calculated as the proportion of SNP heritability accounted for by each module divided by the

proportion of total SNPs within the module. Significance is assessed using a block jack-knife procedure⁴², which accounts for module size and gene length, followed by FDR correction of *P* values.

Data availability statement. Human brain RNA-seq data have been deposited in Synapse (<https://www.synapse.org/#!Synapse:syn4587609>) under accession number syn4587609. Data for the *SOX5* overexpression are available from the Gene Expression Omnibus (accession number GSE89057). All other data are available from the corresponding author upon reasonable request.

Code availability. Code underlying the DGE, differential alternative splicing, cortical patterning, and co-expression network analyses is available at <https://github.com/dhglab/Genome-wide-changes-in-lncRNA-alternative-splicing-and-cortical-patterning-in-autism>.

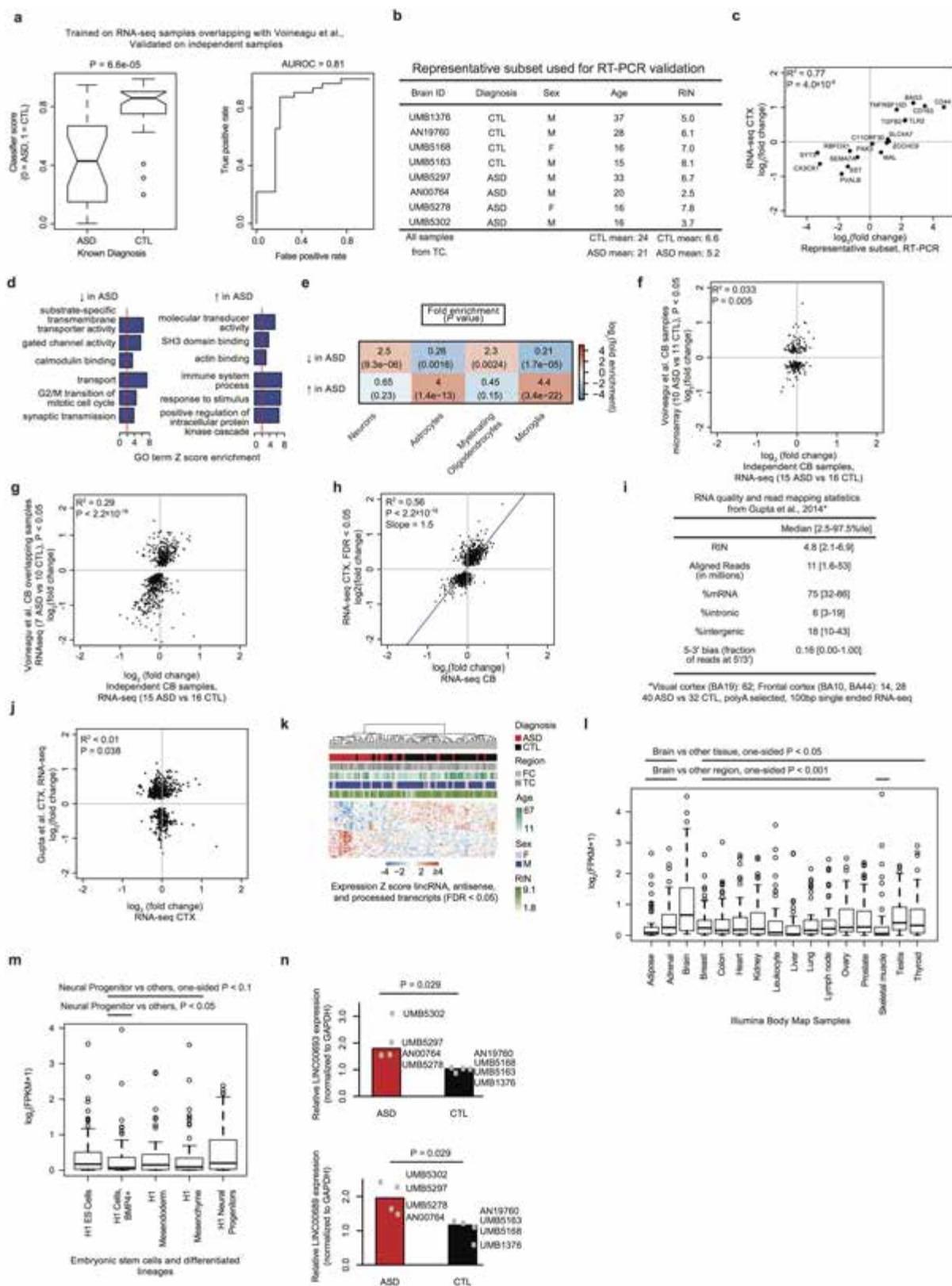
31. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (Suppl. 1), 1–9 (2006).
32. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
33. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
34. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
35. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.* **40**, e61 (2012).
36. Scoles, H. A., Urraca, N., Chadwick, S. W., Reiter, L. T. & Lasalle, J. M. Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. *Mol. Autism* **2**, 19 (2011).
37. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17 (2005).
38. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
39. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1**, 54 (2007).
40. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* **28**, 2209–2210 (2012).
41. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
42. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
43. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
44. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
45. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
47. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
48. Dittmar, K. A. *et al.* Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Mol. Cell. Biol.* **32**, 1468–1482 (2012).
49. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
50. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41**, D996–D1008 (2013).
51. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
52. Langfelder, P. & Horvath, S. Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* **46**, i11 (2012).
53. Winden, K. D. *et al.* The organization of the transcriptional network in specific neuronal classes. *Mol. Syst. Biol.* **5**, 291 (2009).
54. Robinson, E. B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat. Genet.* **48**, 552–555 (2016).
55. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
56. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
57. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
58. Global Lipids Genetics Consortium Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).



Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Methodology, quality control, and differential expression replication analysis. **a**, RNA-seq workflow (see Supplementary Information for details). **b**, RNA-seq quality and alignment statistics from this study, including RNA integrity number (RIN), sequencing depth (aligned reads), proportion of reads mapping to different genomic regions, and bias in coverage from the 5' to the 3' ends of transcripts. **c**, RNA-seq read coverage relative to normalized gene length across transcript length across samples. **d**, Dependence between coverage and RIN across gene body. **e–g**, Correlation of transcript model quantifications comparing the union exon model (used throughout this study), the whole gene model (which includes introns), and the Cufflinks approach⁴³ to estimating FPKM. **h**, Summary table describing the characteristics of the matched covariate data used in the DGE and differential alternative splicing (DS) analysis of ASD in cortex and cerebellum. This includes the number of samples overlapping with our previous work⁸, the age and RIN distributions, and the dependence between diagnosis and age and RIN (summarized from Supplementary Table 1). **i**, Independent replication of ASD versus control DGE fold changes between previously evaluated and new ASD samples in cortex by RNA-seq using samples from ref. 8

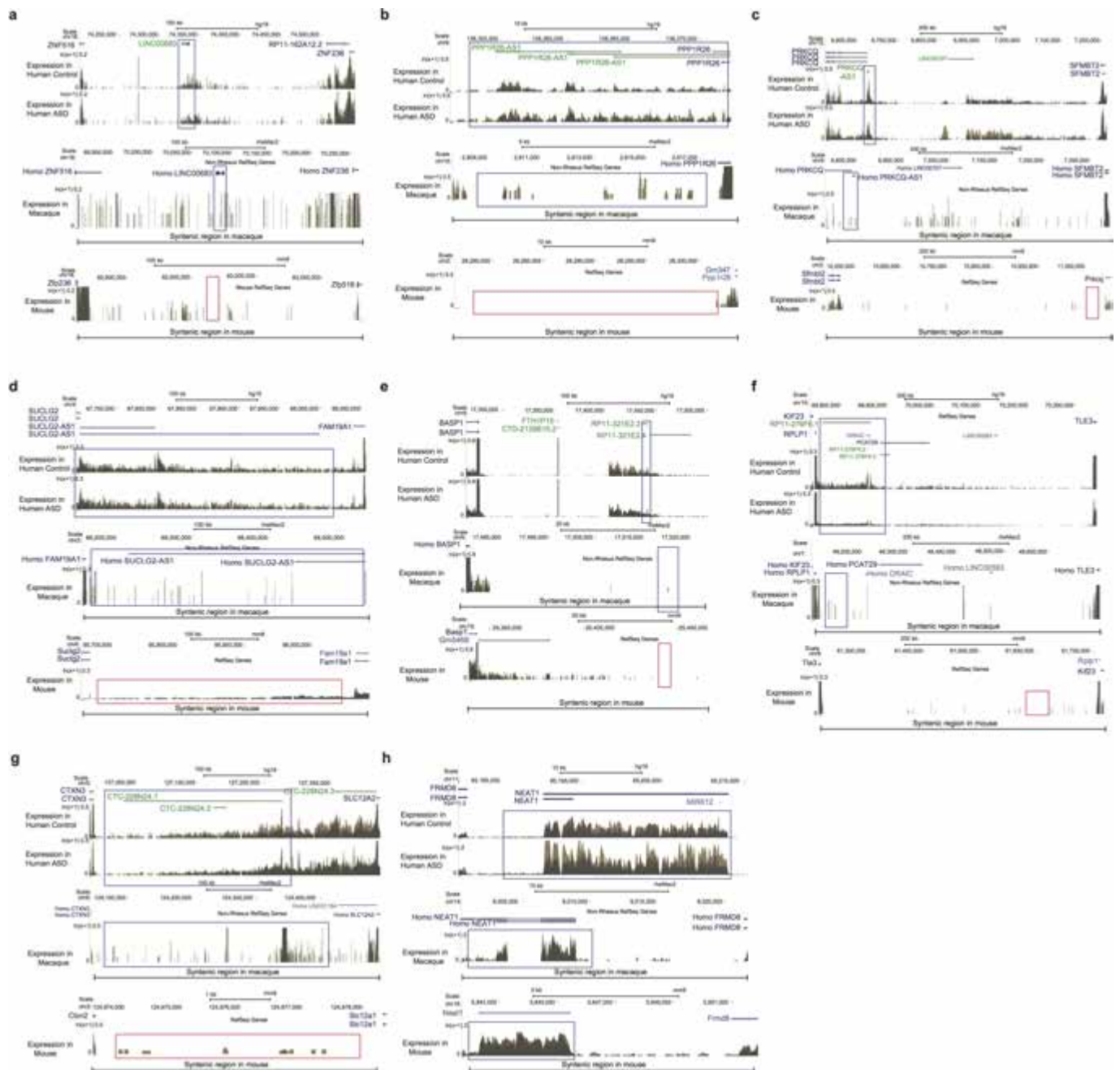
(similar to Fig. 1a, but with RNA-seq in all samples). **j**, Correlation of *P* value rankings with Spearman's correlation across different DGE methods for DGE analysis in cortex, comparing the 'full model' (LME *P* value) described in the Supplementary Information with other methods. Methods include removal of three additional principal components of sequencing surrogate variables (SVs) (LME with 5 SVs, top left), application of a permutation analysis for DGE *P* value computation (LME *P*, permuted, top right), application of variance-weighted linear regression for DGE⁴⁴ (limma voom, middle left), application of surrogate variable analysis for DGE⁴⁵ (full model + 17 SVs, middle right), and application of DESeq2 with the full model⁴⁶, which uses a negative binomial distribution (bottom left). **k**, Comparison of fold changes between frontal cortex (FC) and temporal cortex (TC) for all samples, demonstrating similar changes in both regions. **l**, Average linkage hierarchical clustering of samples in ASD cortex using the top 100 upregulated and top 100 downregulated protein coding genes, demonstrating that confounders do not drive clustering of about two-thirds of samples. **m**, The first principal component of the cortex DGE set is primarily associated with diagnosis, and not with other factors. The red line marks a Bonferroni-corrected $P = 0.05$.



Extended Data Figure 2 | See next page for caption.

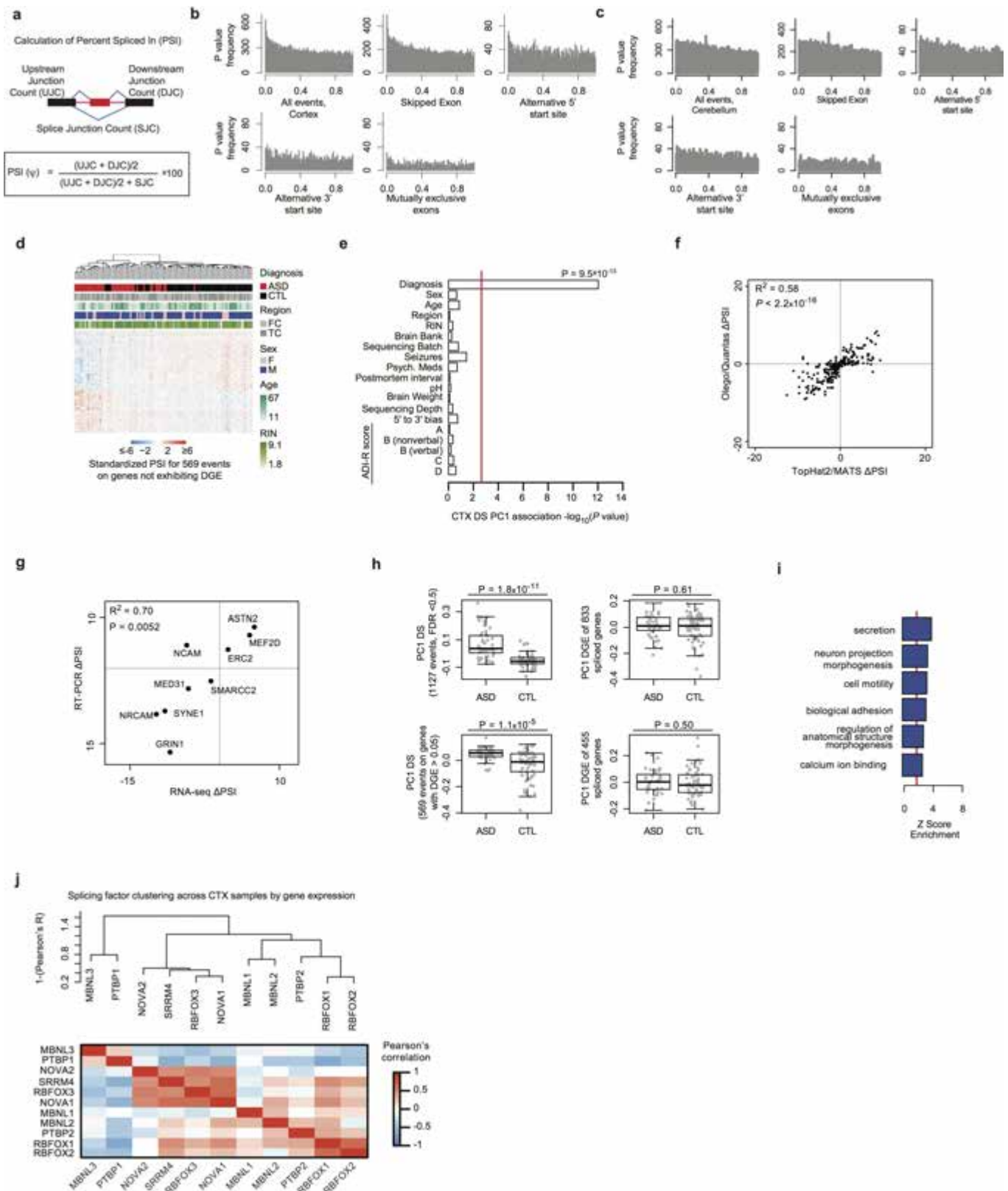
Extended Data Figure 2 | Transcriptome-wide DGE analysis. **a**, We applied a classification method robust to overfitting (elastic net model⁴⁷) by training on the RNA-seq data from samples previously analysed in ref. 8 (Extended Data Fig. 1h, similar to the comparison in Extended Data Fig. 1i) and classifying ASD versus control status in independent samples. Results are shown as a comparison of classification scores (left) and area under the receiver operator characteristic curve (AUROC, right). Approximately 85% of ASD samples are classified successfully around a false positive rate of 20%. **b**, Summary table describing the subset of representative, covariate matched samples used for qRT-PCR validations. Supplementary Table 2 contains the underlying values. **c**, Fold changes from RNA-seq compared against fold changes from qRT-PCR (see Supplementary Table 2 for data). **d**, GO term enrichment analysis of genes that are upregulated or downregulated in individuals with ASD. **e**, Enrichment analysis of cell-type specific gene sets (defined as genes with fivefold higher expression in the cell type than in other cell types) with genes that are decreased or increased in ASD. **f**, **g**, Independent replication analysis of ASD versus control DGE fold changes between previously evaluated and new ASD samples from cerebellum by microarray and RNA-seq using samples from ref. 8 (similar to Fig. 1a and Extended Data Fig. 1i). The RNA-seq data show a replication signal between previously evaluated and new samples from this study. **h**, Comparison of fold changes that were significant at $FDR < 0.05$ in

the ASD versus control DGE analysis from cortex compared with fold changes observed in cerebellum, revealing strong concordance but a lower average fold change in the cerebellum. **i**, Sample summary and quality control (QC) statistics for ref. 4. Compare to Extended Data Fig. 1b and see Supplementary Information for additional discussion. Compared to this study, samples from ref. 4 were prepared by poly(A) selection RNA-seq, exhibit lower RNA integrity number (RIN, median 4.8 versus 7.3), have lower median sequencing depth (11 million versus 40 million), exhibit greater 5'-3' bias, and have generally greater variability across all QC metrics. **j**, Comparison of fold-changes for the top significant genes from ref. 4 ($P < 0.01$ as provided in their Supplementary Information) with the fold changes for the same genes in this study. Co-expression network analysis demonstrated that the moderate agreement is largely driven by concordance in upregulation of microglial genes in both studies (Extended Data Fig. 8e). **k**, Average linkage hierarchical clustering of lncRNAs in the DGE set. **l**, Boxplots of expression values of DGE lncRNAs across multiple tissue types from the Illumina Body Map (expression data from ref. 12). Lines above the plot indicate pairwise significance with a one-sided Wilcoxon rank-sum test between brain and the other tissues. **m**, Similar to **l**, except for embryonic stem cells and stem-cell-derived cell types. **n**, RT-PCR validation of the two lncRNAs shown in Fig. 1c, d; P values computed by two-sided Wilcoxon rank-sum test.



Extended Data Figure 3 | RNA-seq gene expression on genome browser tracks for selected primate-specific lncRNAs in human, macaque and mouse. For each lncRNA, expression for representative samples for ASD versus control (top) in human, macaque (middle), and mouse (bottom) are shown. The genome location for macaque and mouse displayed is

syntenic to the human region, with the expected location of the lncRNA highlighted. **a–g**, Examples of specific lncRNA transcripts that show primate-specific (in human and macaque, or only in human, but not in mouse) expression. **h**, Example of a strongly conserved lncRNA, which shows robust expression in all three species.



Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | Splicing analyses and validation in ASD.

a, Schematic of the PSI metric used for differential alternative splicing³⁵.

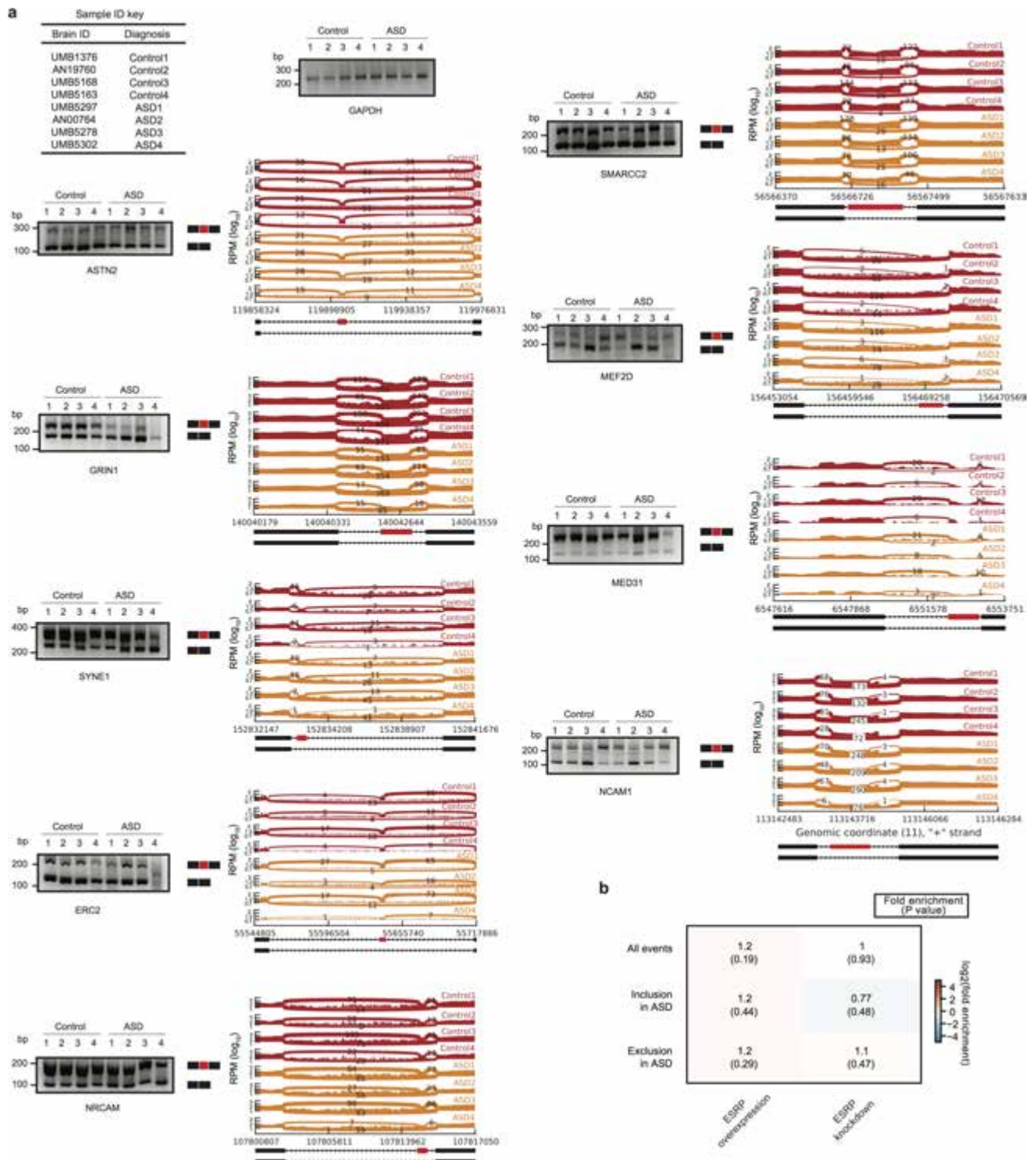
b, Distribution of LME model *P* values for changes in the PSI between ASD and control in cortex for all events and event subtypes. **c**, Distribution of LME model *P* values for changes in the PSI between ASD and control in cerebellum. **d**, Average linkage hierarchical clustering in ASD and control cortex samples using top 100 differentially included and top 100 differentially excluded exons from the differential splicing set.

e, The first principal component of the cortex differential splicing set is strongly associated with diagnosis, but not other factors. Red line marks Bonferroni-corrected $P = 0.05$. **f**, Comparison of the cortex differential splicing with the pipeline used here (TopHat2 (ref. 43) followed by multivariate analysis of transcript splicing, MATS³⁵) with PSI values obtained via another method (read alignment by OLEgo followed by PSI

quantification with Quantas¹⁵). **g**, Comparison of Δ PSI values between RT-PCR and RNA-seq for nine splicing events (Supplementary Table 3).

h, Differential splicing analysis identifies events independent of DGE signal. Top, difference between ASD and control in the differential splicing set based on PC1 of the differential splicing set at the PSI level, and PC1 of the gene expression levels of genes in the differential splicing set. Bottom, same comparison after removing nominally differentially expressed genes ($P < 0.05$). *P* values computed by two-sided Wilcoxon rank-sum test.

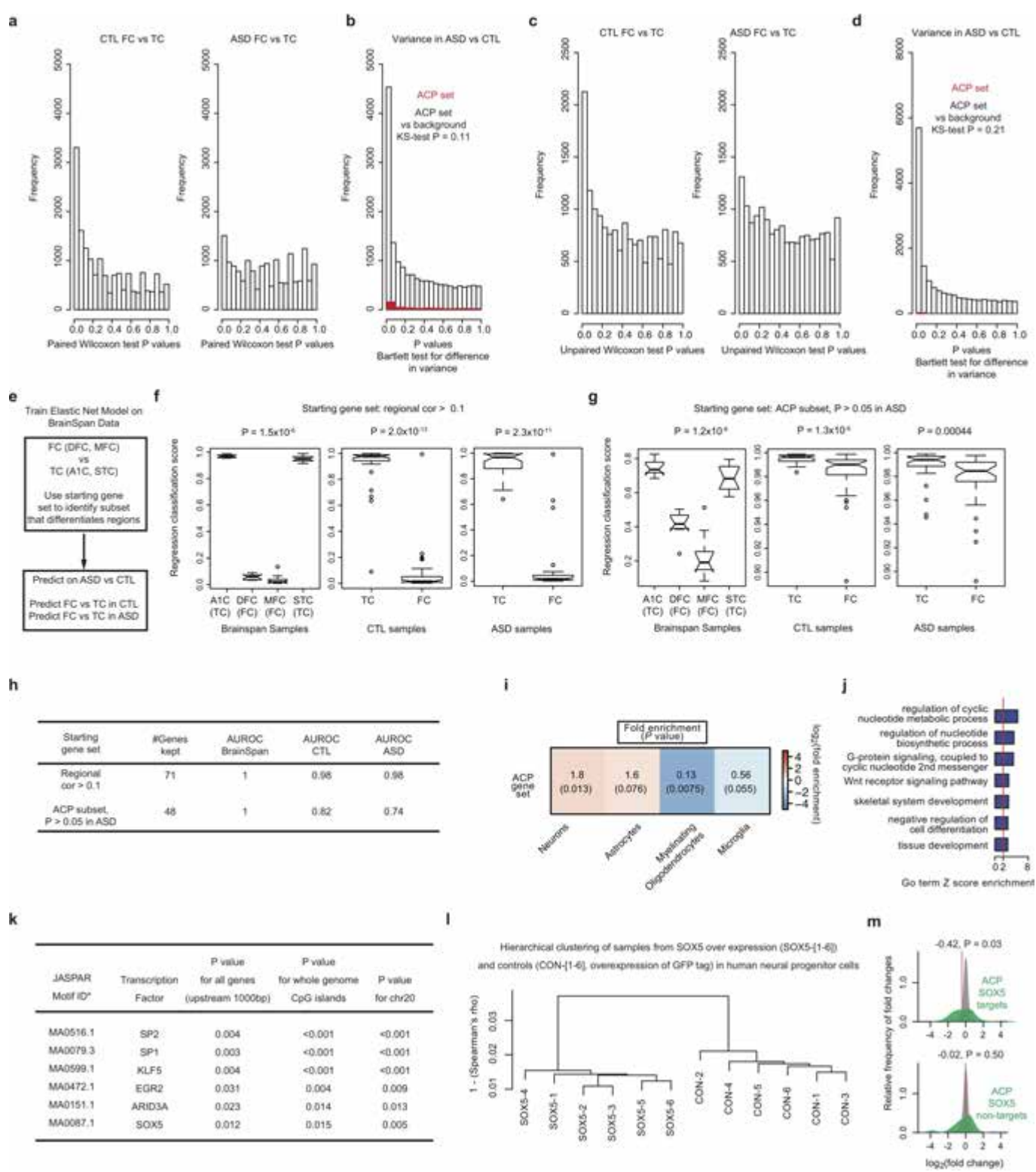
i, GO term enrichment analysis of genes with differential splicing events in ASD. **j**, Clustering dendrogram and heat map for neuronal splicing factor gene expression levels across samples demonstrating three major clusters and the known positive correlation between *SRRM4* and *RBFOX1* and anticorrelation between *PTBP1* and *SRRM4* (refs 14,19).



Extended Data Figure 5 | Additional splicing analyses in ASD.

a, PCR validation and sashimi plots for nine splicing events delineated in Extended Data Fig. 4d, from the indicated samples (see Extended Data Fig. 2b for details of these samples). Notably, these genes are not in the DGE set, but are detected in the differential alternative splicing set owing to altered transcript structure. **b**, Heat map as in Fig. 1h for the

splicing regulator *ESRP*⁴⁸. *ESRP* is not known to be involved in neuronal function, *ESRP1* is not expressed in cortex, and *ESRP2* is expressed but not significantly different between ASD and control cortex. Therefore, we show *ESRP* enrichment analysis in differential splicing events as a control for Fig. 1h. Enrichment *P* values are computed as described in Methods.

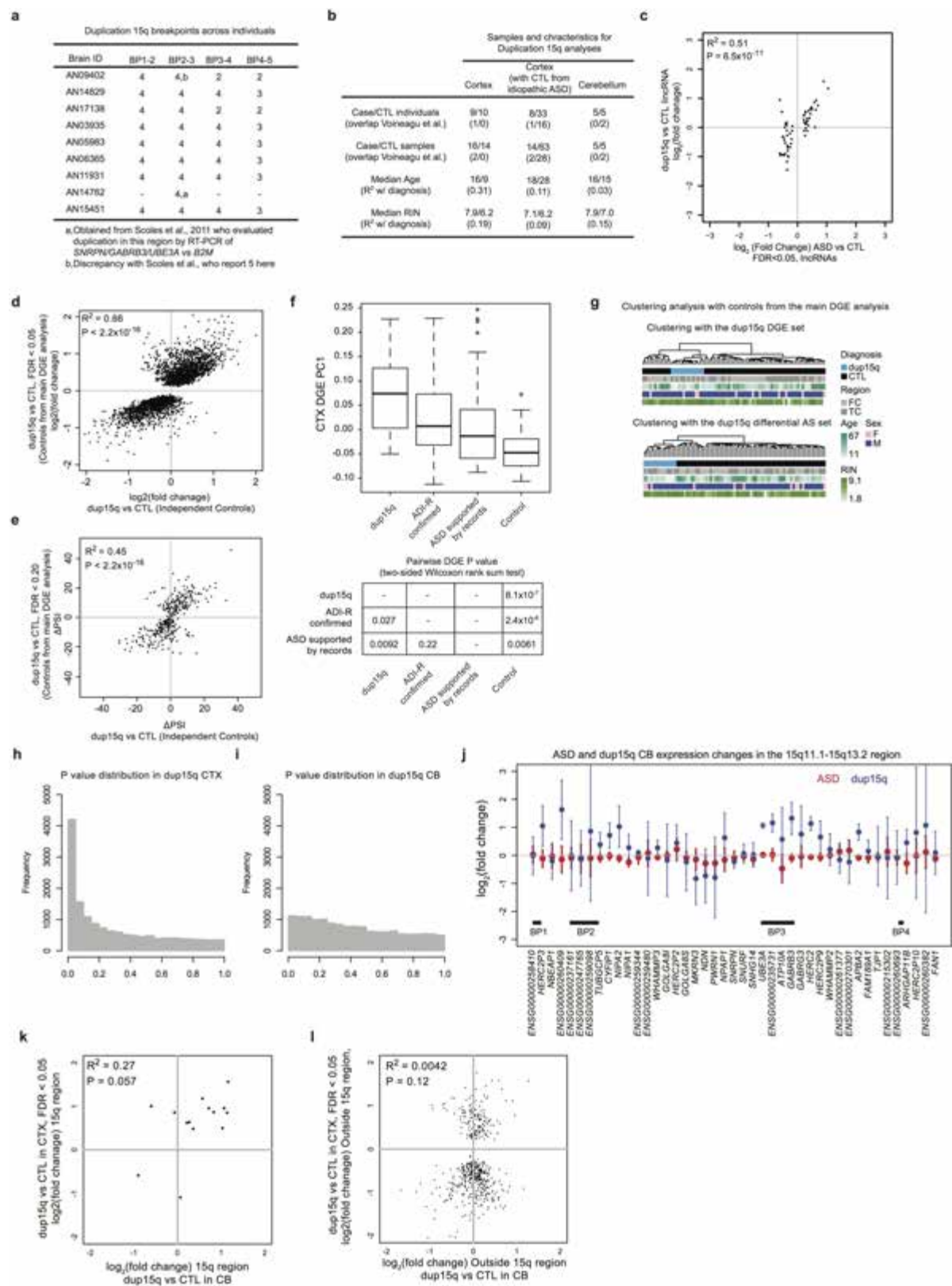


Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Attenuation of cortical patterning in ASD.

a, Histograms of P values from paired Wilcoxon rank-sum test differential gene expression between 16 frontal cortex (FC) and 16 temporal cortex (TC) samples from control and ASD individuals. **b**, Histogram of Bartlett's test P values for differences in gene expression variance between ASD and control samples for all genes (white) and genes in the ACP set (red). The Kolmogorov–Smirnov ($K-S$) test P value for a difference between these two distributions is shown. **c**, Histograms of P values from unpaired Wilcoxon rank-sum test DGE between 21 frontal cortex and 22 temporal cortex samples after removing those used in ref. 8. **d**, Histogram of Bartlett's test P values for differences in gene expression variance between ASD and control samples for all genes (white) and genes in the ACP set (red). The Kolmogorov–Smirnov test P value for a difference between these two distributions is reported. **e**, Approach to training the elastic net model on BrainSpan^{49,50} frontal cortex and temporal cortex samples and application of the model to 123 cortical samples in this study. **f–h**, Results of learned cortical region classifications with different starting gene sets,

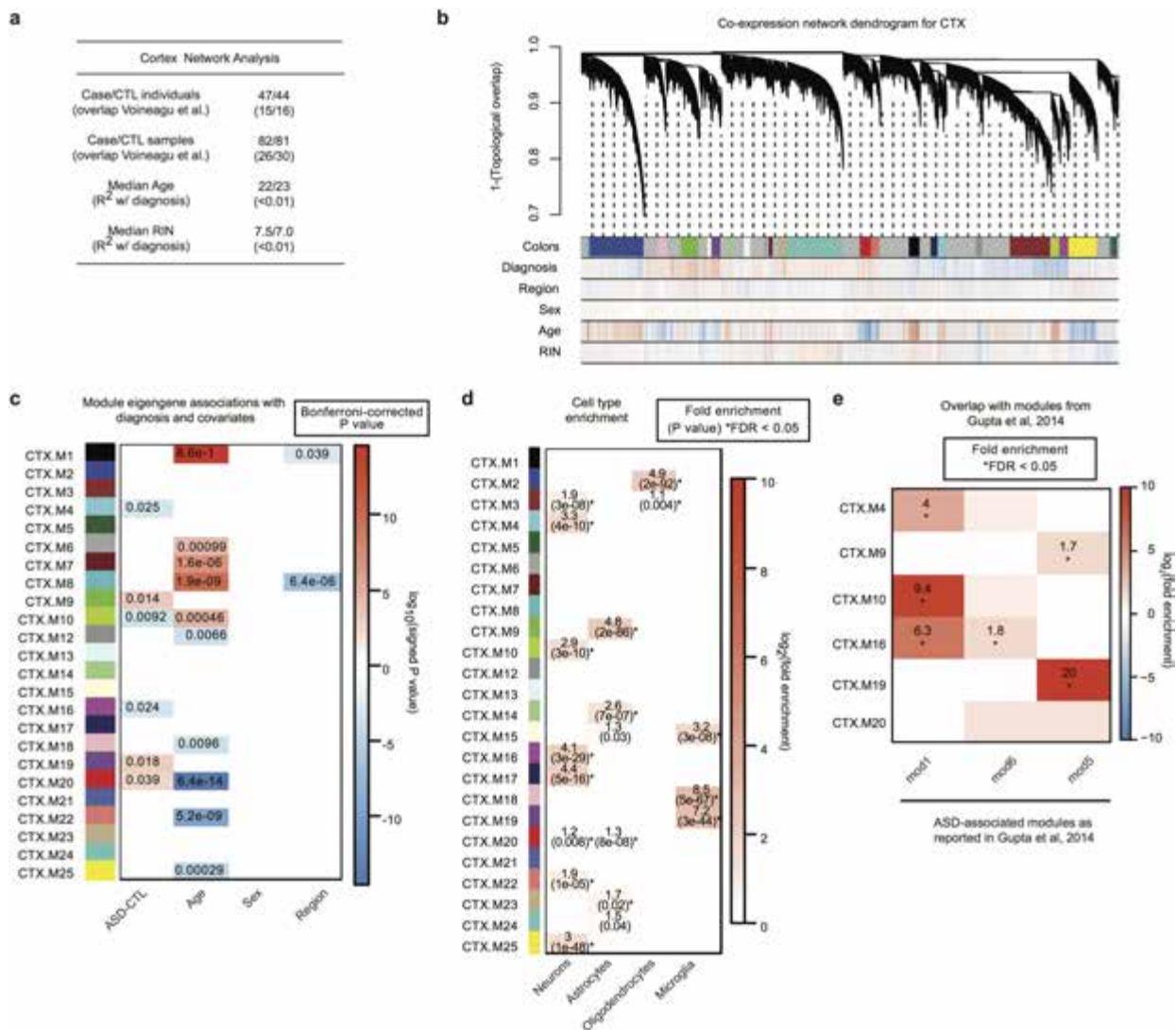
with the BrainSpan training set (left), control samples (middle) and ASD samples (right) in each panel and the Wilcoxon rank-sum test P value of frontal versus temporal cortex difference for each comparison. A1C, primary auditory cortex; DFC, dorsolateral prefrontal cortex; MFC, medial prefrontal cortex; STC, superior temporal cortex. **i**, Cell-type enrichment analysis for genes in the ACP set. **j**, GO term enrichment analysis of the ACP set. Enrichment P values are computed as described in Methods. **k**, Enrichment statistics for transcription factor motifs found to be significantly enriched in the ACP set (see Supplementary Information for details of P value computation). **l**, Average linkage hierarchical clustering of the global gene expression profiles for samples with overexpression of SOX5 and green fluorescent protein (GFP) tag overexpression (controls). **m**, Density plots of fold changes for the subset of ACP genes that are predicted SOX5 targets (top, green) and non-targets (bottom, green) against background (grey). The median \log_2 [fold change] is marked (red line) and P values are from a one-sided Wilcoxon rank-sum test.



Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Duplication 15q syndrome analyses. **a**, Copy number between breakpoints in the 15q region. Genome-wide copy number analysis allowed evaluation of copy number in additional regions from previous studies³⁶. **b**, Sample characteristics for the dup15q analyses (additional details available in Supplementary Table 1). **c**, Similar to Fig. 3b, but focusing on the lncRNAs found to be significantly differentially expressed in idiopathic ASD compared to control subjects. **d**, Comparison of DGE fold changes demonstrating that using different control samples (control samples used in the idiopathic analysis, column 2 of Extended Data Fig. 7b) for the dup15q cortex analysis yields similar findings. **e**, Similar to **d** except for the differential alternative splicing analysis. **f**, Comparison of heterogeneity in the DGE signal using the first principal component of the ASD cortex DGE set across all cortical samples used

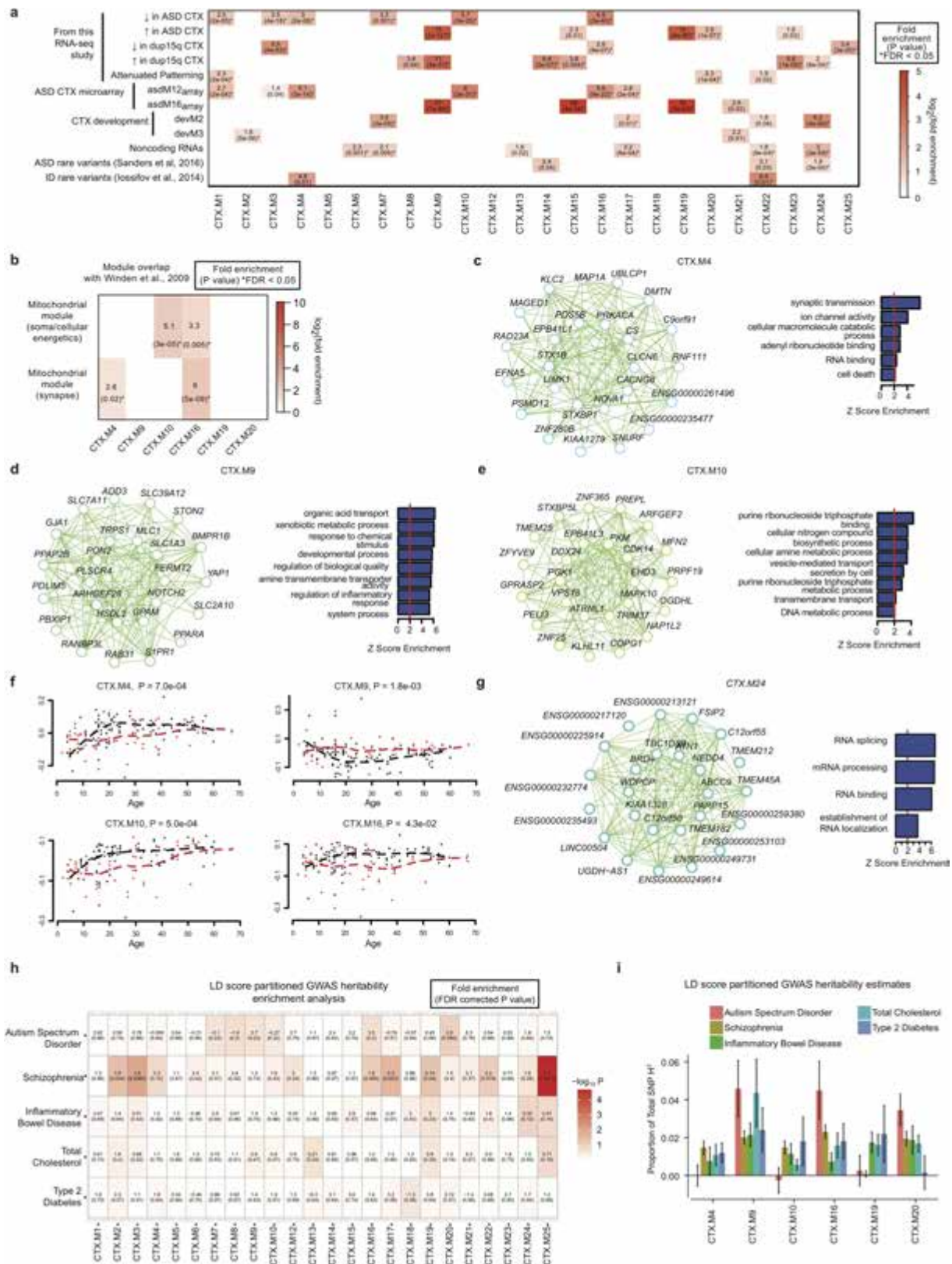
in DGE analyses. Samples from individuals with diagnoses confirmed by dup15q mutations, confirmed by Autism Diagnostic Interview-Revised (ADI-R), and supported by clinical records are all significantly different from controls by two-sided pairwise Wilcoxon rank sum tests. **g**, Similar to Fig. 3d, but with the larger set of controls from the idiopathic ASD versus control analysis in Fig. 1. **h**, **i**, *P* value distributions for DGE changes outside the 15q region for cortex and cerebellum. **j**, Similar to Fig. 3a, but for the cerebellum analysis. **k**, Comparison of significant DGE changes in the duplicated region from cortex with changes in cerebellum. **l**, Comparison of significant DGE changes outside of the dup15q region in cortex with changes in cerebellum. Scatter plot *P* values correspond to the statistical significance of the Pearson correlation coefficient between fold changes (see Methods).



Extended Data Figure 8 | Cortex co-expression network analyses.

a, Sample characteristics for the cortex network analyses; additional details available in Supplementary Table 1. **b**, Average linkage hierarchical clustering using the topological overlap metric for co-expression dissimilarity³⁷. Modules are identified from this dendrogram, which was constructed from a consensus of 100 bootstrapped datasets^{51,52} (see Methods). Correlations for each gene to covariates are delineated below the dendrogram (blue, negative; red, positive). Modules are labelled with colours and numerical labels (see Supplementary Table 4

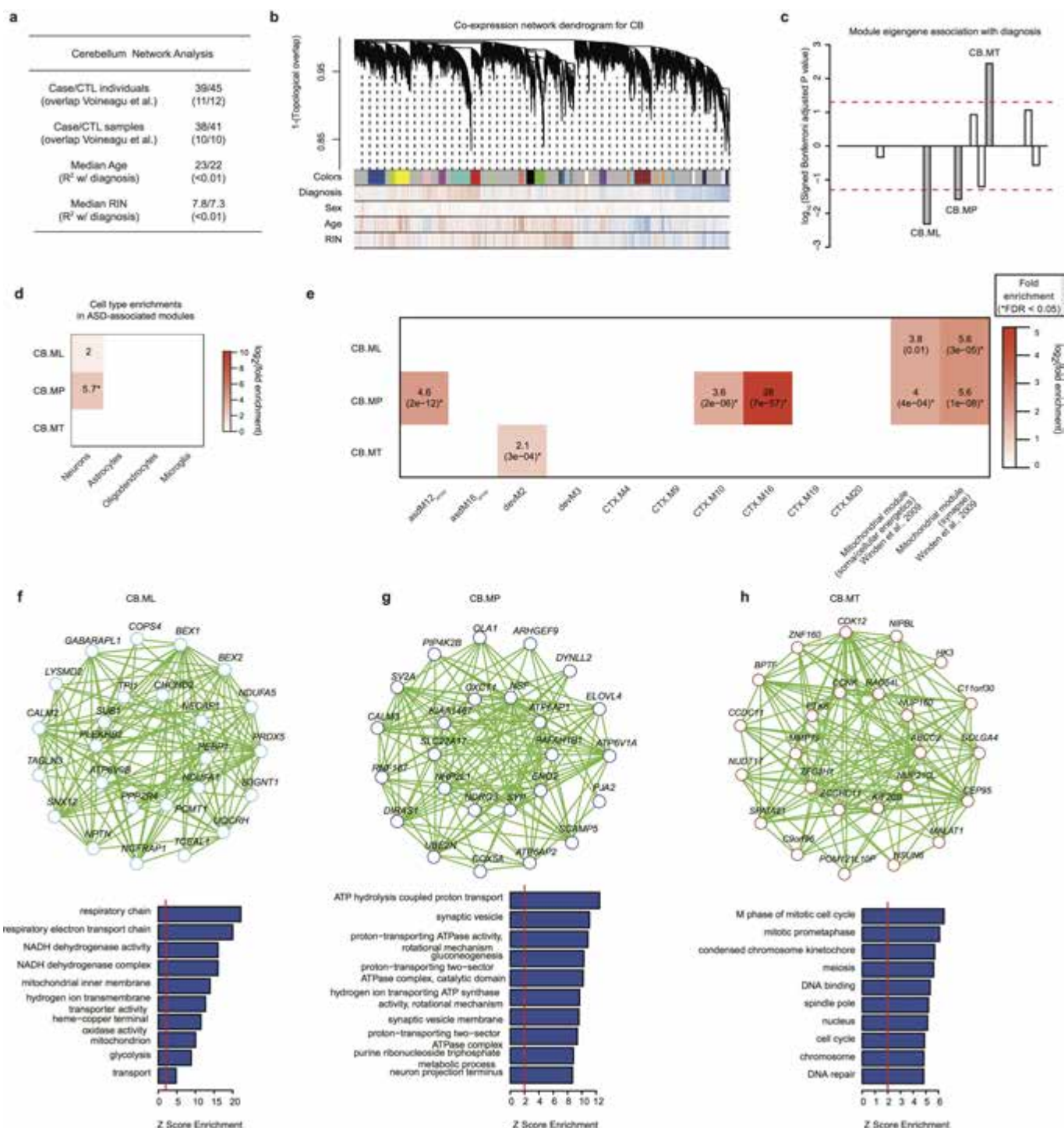
for additional details). CTX.M11 is a module of genes that are not co-expressed (grey module) and was not evaluated in further comparisons. **c**, Module-trait associations as computed by an LME model with all factors on the x axis used as covariates. Technical covariates were removed as part of adjusting the FPKM values. All *P* values are displayed where the association passed Bonferroni-corrected $P < 0.05$. **d**, Module enrichments for cell-type specific gene expression patterns. Asterisks indicate $FDR < 0.05$ across all comparisons. **e**, Enrichment of ASD-associated modules with that from ref. 4. * $FDR < 0.05$ (see Supplementary Table 4 for details).



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Additional figures for cortex co-expression network analyses. **a**, Gene set enrichment analyses comparing the 24 cortex co-expression modules with multiple gene sets from this RNA-seq study, post-mortem ASD cortex microarray⁸, human cortical development¹⁰, the set of all brain-expressed lncRNAs, genes enriched for ASD-associated rare variants²⁶, and genes with *de novo* variants associated with intellectual disability (ID)⁹. Boxes are filled if the odds ratio is greater than 0 and the enrichment $P < 0.05$. *FDR < 0.05 across all comparisons, controlling for gene length and expression level with logistic regression (Supplementary Information). **b**, Overlap of gene sets between firing-rate and mitochondrial associated modules from ref. 53 with ASD-associated modules in cortex. **c–e**, Module plot of ASD-associated modules not shown in Fig. 4 (CTX.M4, CTX.M9, CTX.M10) displaying the top

hub genes along with the module's GO term enrichment. **f**, Temporal trajectories for four module eigengenes (CTX.M4, CTX.M9, CTX.M10, CTX.M16) associated with ASD, similar to Fig. 4g. ASD samples are represented by red points and lines, control samples by black. **g**, Module plot and GO term enrichment for CTX.M24, which is enriched in ASD-associated rare variants and lncRNAs. **h**, Common variant enrichment across modules as calculated by GWAS enrichment with LD score regression^{41,42} (see Methods). Disease GWAS studies evaluated include ASD⁵⁴, schizophrenia⁵⁵, inflammatory bowel disease⁵⁶, type 2 diabetes mellitus⁵⁷ and serum lipid levels⁵⁸. P values are FDR corrected across all GWAS studies and modules. **i**, Plot of the proportion of SNP heritability across diseases for ASD-associated modules. Error bars represent s.e.



Extended Data Figure 10 | Cerebellum co-expression network analyses.

a, Sample characteristics for the cerebellum network analyses; additional details available in Supplementary Table 1. **b**, Modules identified from a dendrogram constructed from a consensus of 100 bootstrapped networks (see Methods). Correlations for each gene to each measured factor are delineated below the dendrogram (blue, negative; red, positive). Modules are labelled alphabetically instead of numerically to distinguish them from the cortex modules. Additional information is available in Supplementary Table 4. **c**, Signed association of module eigengenes with diagnosis; positive values indicate modules with increased expression in ASD samples. Grey bars with labels signify three ASD-associated

modules. **d**, Cell-type enrichments for the three ASD-associated modules. **e**, Gene set enrichment analyses comparing the three ASD-associated cerebellum modules with post-mortem ASD cortex microarray, human brain development, six cortex ASD-associated modules from this RNA-seq study, and firing rate and mitochondrial associated modules from ref. 53. Boxes are filled if the odds ratio is greater than 0 and the enrichment $P < 0.05$. *FDR < 0.05 across all comparisons. **f–h**, Module plots of CB.ML, CB.MP, and CB.MT displaying the top hub genes along with the GO term enrichment. Additional details, including module preservation statistics for cerebellum in cortex and vice versa, are available in Supplementary Table 4.

Epigenetic stress responses induce muscle stem-cell ageing by *Hoxa9* developmental signals

Simon Schwörer¹, Friedrich Becker¹, Christian Feller², Ali H. Baig¹, Ute Köber¹, Henriette Henze¹, Johann M. Kraus³, Beibei Xin⁴, André Lechel⁵, Daniel B. Lipka⁶, Christy S. Varghese¹, Manuel Schmidt¹, Remo Rohs⁴, Ruedi Aebersold^{2,7}, Kay L. Medina⁸, Hans A. Kestler^{1,3}, Francesco Neri¹, Julia von Maltzahn^{1,§}, Stefan Tümpel^{1,§} & K. Lenhard Rudolph^{1,9}

The functionality of stem cells declines during ageing, and this decline contributes to ageing-associated impairments in tissue regeneration and function¹. Alterations in developmental pathways have been associated with declines in stem-cell function during ageing^{2–6}, but the nature of this process remains poorly understood. Hox genes are key regulators of stem cells and tissue patterning during embryogenesis with an unknown role in ageing^{7,8}. Here we show that the epigenetic stress response in muscle stem cells (also known as satellite cells) differs between aged and young mice. The alteration includes aberrant global and site-specific induction of active chromatin marks in activated satellite cells from aged mice, resulting in the specific induction of *Hoxa9* but not other Hox genes. *Hoxa9* in turn activates several developmental pathways and represents a decisive factor that separates satellite cell gene expression in aged mice from that in young mice. The activated pathways include most of the currently known inhibitors of satellite cell function in ageing muscle, including Wnt, TGF β , JAK/STAT and senescence signalling^{2–4,6}. Inhibition of aberrant chromatin activation or deletion of *Hoxa9* improves satellite cell function and muscle regeneration in aged mice, whereas overexpression of *Hoxa9* mimics ageing-associated defects in satellite cells from young mice, which can be rescued by the inhibition of *Hoxa9*-targeted developmental pathways. Together, these data delineate an altered epigenetic stress response in activated satellite cells from aged mice, which limits satellite cell function and muscle regeneration by *Hoxa9*-dependent activation of developmental pathways.

Age-dependent declines in the number and function of Pax7⁺ satellite cells (SCs) impair the regenerative capacity of skeletal muscle^{2,4,9}. Genes and pathways that contribute to this process^{2–6} often also have a role in regulating embryonic development^{10–13}. Despite these parallels, the function of the master regulators of development, Hox genes, has not been determined in SC ageing. An analysis of freshly isolated, *in vivo* activated SCs from young adult and aged mice (Extended Data Fig. 1a–e) revealed a specific upregulation of *Hoxa9* in SCs from aged mice, both at the mRNA (Fig. 1a, Extended Data Fig. 2a, b) and protein level (Fig. 1b, c). Similar results were obtained by immunofluorescence staining of SCs (Extended Data Fig. 2c) and myofibre-associated SCs (Fig. 1d, e, Extended Data Fig. 2d) that were activated in culture (Extended Data Fig. 1f, g).

Ageing reduces the proliferative and self-renewal capacity of SCs in wild-type mice^{2,9,14,15} (*Hoxa9*^{+/+}; Extended Data Fig. 3). Homozygous deletion of *Hoxa9* (*Hoxa9*^{−/−}) did not affect the colony-forming capacity of SCs from young adult mice but ameliorated ageing-associated impairment in colony formation of single-cell-sorted SCs in culture (Fig. 2a). *Hoxa9* deletion also increased the self-renewal of myofibre-associated SCs from aged mice in culture but had no effect on SCs

from young adult mice under these conditions (Extended Data Fig. 4a–c). Similar results were obtained by short interfering RNA (siRNA)-mediated knockdown of *Hoxa9* in myofibre-associated SC cultures derived from aged mice (Extended Data Fig. 4d–h). The number of SCs decreases in resting tibialis anterior muscle of ageing wild-type mice^{2,4,9}; this phenotype was not affected by *Hoxa9* gene status (Extended Data Fig. 5a). However, homozygous deletion or siRNA-mediated knockdown of *Hoxa9* increased the total number of Pax7⁺ SCs (Fig. 2b, Extended Data Fig. 5b–e) and improved myofibre regeneration

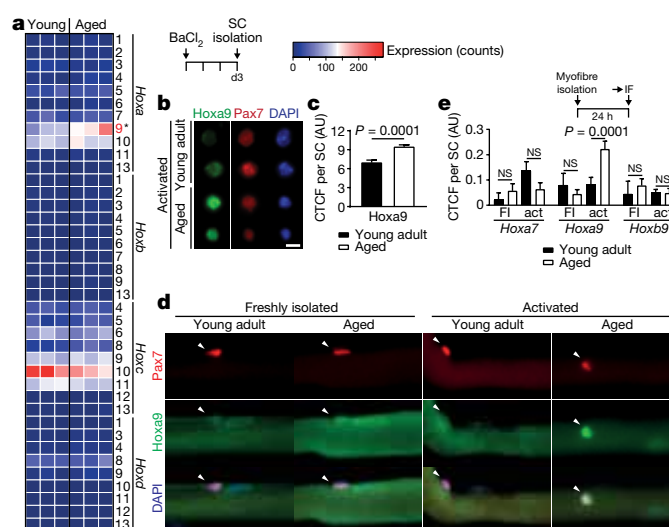


Figure 1 | Upregulation of *Hoxa9* in aged activated SCs. **a–c**, Analysis of freshly isolated, *in vivo* activated SCs (3 days after muscle injury with BaCl₂) from young adult and aged mice. **a**, Heatmap showing the mRNA expression of all Hox genes as determined by RNA-sequencing analysis. **b**, Representative immunofluorescence staining for Hoxa9 and Pax7. Nuclei were counterstained with 4',6-diamidino-2-phenylindole (DAPI). **c**, Corrected total cell fluorescence (CTCF) for Hoxa9 per SC as shown in **b**. AU, arbitrary units. **d**, **e**, Immunofluorescence (IF) staining for Hoxa9 and Pax7 in myofibre-associated SCs that were quiescent (freshly isolated (FI) myofibres) or activated (act; 24 h culture of myofibres). **d**, Representative images with arrowheads denoting Pax7⁺ cells. **e**, CTCF for indicated Hox genes. Note the specific induction of *Hoxa9* in activated SCs isolated from aged mice. Scale bars, 5 μ m (**b**) and 20 μ m (**d**). *P* values were calculated by two-sided Mann–Whitney *U*-test (**c**) or two-way analysis of variance (ANOVA) (**e**). NS, not significant. *n* = 3 mice in **a**; *n* = 134 nuclei (young), *n* = 181 nuclei (aged) from 3 mice in **c**; *n* = 12/13/17/56 nuclei (*Hoxa7*), *n* = 9/42/102/62 nuclei (*Hoxa9*), *n* = 7/35/34/25 nuclei (*Hoxb9*) from 2 young and 4 aged mice in **e**.

¹Leibniz-Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena, Germany. ²Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, Auguste-Piccard-Hof 1, 8093 Zürich, Switzerland. ³Institute of Medical Systems Biology, Ulm University, James-Frank-Ring, 89081 Ulm, Germany. ⁴Molecular and Computational Biology Program, University of Southern California, 1050 Childs Way, Los Angeles, California 90089, USA. ⁵Department of Internal Medicine I, Ulm University, Albert-Einstein-Allee 23, 89081 Ulm, Germany. ⁶Division of Epigenomics and Cancer Risk Factors, DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. ⁷Faculty of Science, University of Zürich, Zürich, Switzerland. ⁸Department of Immunology, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905, USA. ⁹Faculty of Medicine, Friedrich-Schiller-University, Jena, Germany. §These authors jointly supervised this work.

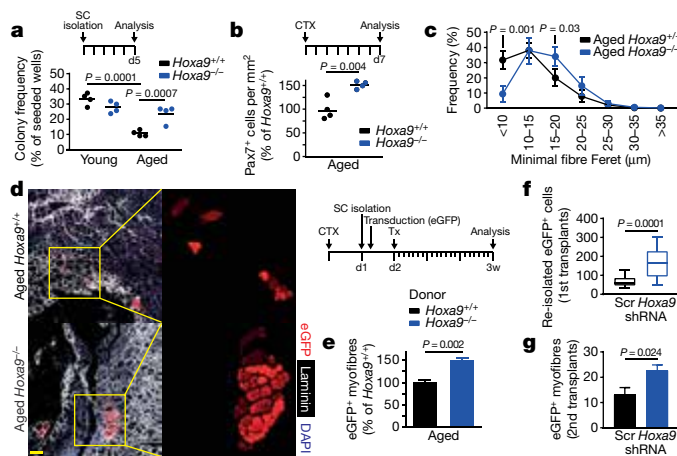


Figure 2 | *Hoxa9* deficiency improves muscle regeneration in aged mice.

a, Frequency of myogenic colonies derived from single-cell-sorted SCs from young adult or aged *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice after 5 days (d5) of culture. **b**, **c**, Quantification of Pax7⁺ cells per area (**b**) and frequency distribution of minimal Feret's diameter (**c**) of tibialis anterior muscle fibres from aged *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice, 7 days after muscle injury with cardiotoxin (CTX). **d**, **e**, Transplantation (Tx) of enhanced green fluorescent protein (eGFP)-labelled SCs from aged *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice. **d**, Representative immunofluorescence staining for eGFP, laminin and DAPI in engrafted tibialis anterior muscles. Scale bar, 50 μ m. **e**, Quantification of donor-derived (eGFP⁺) myofibers in **d**. **f**, Quantification of donor-derived (eGFP⁺) SCs re-isolated from primary recipients. Scr, scrambled control shRNA. **g**, Quantification of donor-derived (eGFP⁺) myofibers from secondary recipients. Data in **f** represent median with 50% confidence interval box and 95% confidence interval whiskers. *P* values were calculated by two-way ANOVA (**a**, **c**), two-sided Student's *t*-test (**b**, **e**, **g**), or two-sided Mann-Whitney *U*-test (**f**). *n* = 4 mice in **a**; *n* = 4 mice in **b**, **c**; *n* = 8 recipient mice in **e**; *n* = 20 recipient mice in **f**; *n* = 5 recipient mice in **g**.

in injured muscle of aged mice almost to the levels in young adult mice (Fig. 2c, Extended Data Fig. 5f), albeit without affecting overall SC proliferation rates seven days after muscle injury (Extended Data Fig. 5g, h). *Hoxa9* gene deletion also improved the cell-autonomous, *in vivo* regenerative capacity of transplanted SCs derived from aged donor mice but did not affect the capacity of SCs derived from young adult donors (Fig. 2d, e, Extended Data Fig. 6a). Similarly, *Hoxa9* downregulation by short hairpin RNA (shRNA) infection rescued the regenerative capacity and the engraftment of transplanted SCs derived from aged mice almost to the level of SCs from young adult mice (Extended Data Fig. 6b–h). When transduced at similar infection efficiency (Extended Data Fig. 6i), *Hoxa9* shRNA compared to scrambled shRNA improved the self-renewal of serially transplanted SCs from aged mice in primary recipients (Fig. 2f, Extended Data Fig. 6j) as well as the regenerative capacity of 500 re-isolated SCs from primary donors that were transplanted for a second round into the injured tibialis anterior muscle of secondary recipients (Fig. 2g, Extended Data Fig. 6k). Together, these results demonstrate that the induction of *Hoxa9* limits SC self-renewal and muscle regeneration in aged mice, and that the deletion of *Hoxa9* is sufficient to revert these ageing-associated deficiencies.

The expression of *Hoxa9* in development and leukaemia is actively maintained by Mll1-dependent tri-methylation at lysine 4 of histone 3 (H3K4me3)^{16–18}. Chromatin immunoprecipitation (ChIP) revealed that H3K4me3 is strongly enriched at the promoter and first exon of *Hoxa9* in activated SCs from aged compared to young adult mice, which was not detected to the same extent for other *Hoxa* genes (Fig. 3a, Extended Data Fig. 7a). ChIP analyses for Mll1 and Wdr5 (a scaffold protein of the Mll1 complex) revealed increased recruitment of these factors to the *Hoxa* cluster with Wdr5 enrichment being confined to the *Hoxa9* locus (Fig. 3b, c). Although no changes were

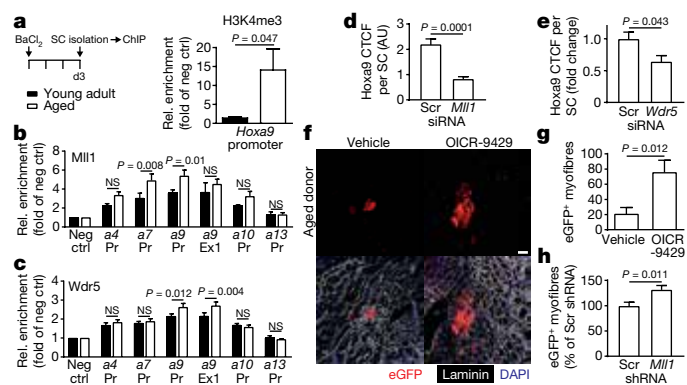


Figure 3 | Mll1 complex-dependent chromatin modification induces *Hoxa9* and limits muscle regeneration in aged mice.

a–c, ChIP-quantitative PCR (qPCR) analysis of the indicated *Hox* promoters (Pr) and exons (Ex) in activated SCs from young adult and aged mice using antibodies against H3K4me3 (**a**), Mll1 (**b**), or Wdr5 (**c**). **d**, **e**, CTFC for *Hoxa9* per SC after *Mll1* siRNA (**d**) or *Wdr5* siRNA transfection (**e**) of freshly isolated myofibre-associated SCs from aged mice. **f–h**, Transplantation of eGFP-labelled SCs from aged mice. **f**, Representative immunofluorescence staining for eGFP, laminin and DAPI in engrafted tibialis anterior muscles after transplantation of OICR-9429 treated SCs. Scale bar, 50 μ m. **g**, **h**, Quantification of donor-derived (eGFP⁺) myofibers after transplantation of OICR-9429-treated (**g**) or shRNA-treated (**h**) SCs. *P* values were calculated by two-way ANOVA (**b**, **c**), two-sided Student's *t*-test (**a**, **g**, **h**) or two-sided Mann-Whitney *U*-test (**d**, **e**). *n* = 6 mice in **a**; *n* = 7 mice (young), *n* = 10 mice (aged) in **b**, **c**; *n* = 109 nuclei (Scr siRNA), *n* = 110 nuclei (*Mll1* siRNA) from 3 mice in **d**; *n* = 116 nuclei (Scr siRNA), *n* = 65 nuclei (*Wdr5* siRNA) from 3 mice in **e**; *n* = 5 recipient mice in **g**; *n* = 6 recipient mice in **h**.

observed for Mll1, both H3K4me3 and Wdr5 showed significantly increased levels in nuclei of myofibre-associated SCs from aged versus young adult mice upon activation (Extended Data Fig. 7b–e). Of note, knockdown of either *Mll1* (also known as *Kmt2a*) or *Wdr5* reduced H3K4me3 levels as well as Mll1 recruitment to the *Hoxa9* locus and ameliorated *Hoxa9* induction in activated myofibre-associated SCs from aged mice (Fig. 3d, e, Extended Data Fig. 7f–i). Similar results were obtained by treatment of aged myofibre-associated SCs with OICR-9429, an inhibitor of the Mll1–Wdr5 interaction¹⁹ (Extended Data Fig. 7j, k). Moreover, both *Mll1* knockdown and OICR-9429 treatment increased the self-renewal and lowered the myogenic commitment of myofibre-associated SCs from aged mice (Extended Data Fig. 7l–q), resulting in increased SC numbers in cultures of purified SCs or myofibre-associated SCs derived from aged mice (Extended Data Fig. 7r, s). Notably, Mll1 inhibition by either stable shRNA knockdown (Extended Data Fig. 7t) or OICR-9429 treatment improved the regenerative capacity of SCs from aged mice when transplanted into injured muscle of recipient mice (Fig. 3f–h). Taken together, these experiments demonstrate that the Mll1 complex contributes to *Hoxa9* induction in activated SCs from aged mice, resulting in impairment in SC function and muscle regeneration. Pax7 expression was downregulated in activated SCs of aged mice (Extended Data Fig. 7u–w) and did not correlate with *Hoxa9* expression (Extended Data Fig. 7x, y), indicating that Mll1-dependent regulation of Pax7 target genes²⁰ was not involved in the Mll1-dependent induction of *Hoxa9* in activated SCs from aged mice.

Next, a global analysis of histone post-translational modifications was carried out on freshly isolated SCs obtained before muscle injury (quiescent state) or two, three and five days after *in vivo* SC activation mediated by muscle injury (Fig. 4a, b, Extended Data Fig. 8a). Using a recently developed mass-spectrometry-based proteomic strategy²¹, 46 histone H3 and H4 lysine acetylation and methylation motifs were quantified. Quiescent SCs from aged mice compared to young adult mice showed increased levels of repressive marks (H3K9me2 and

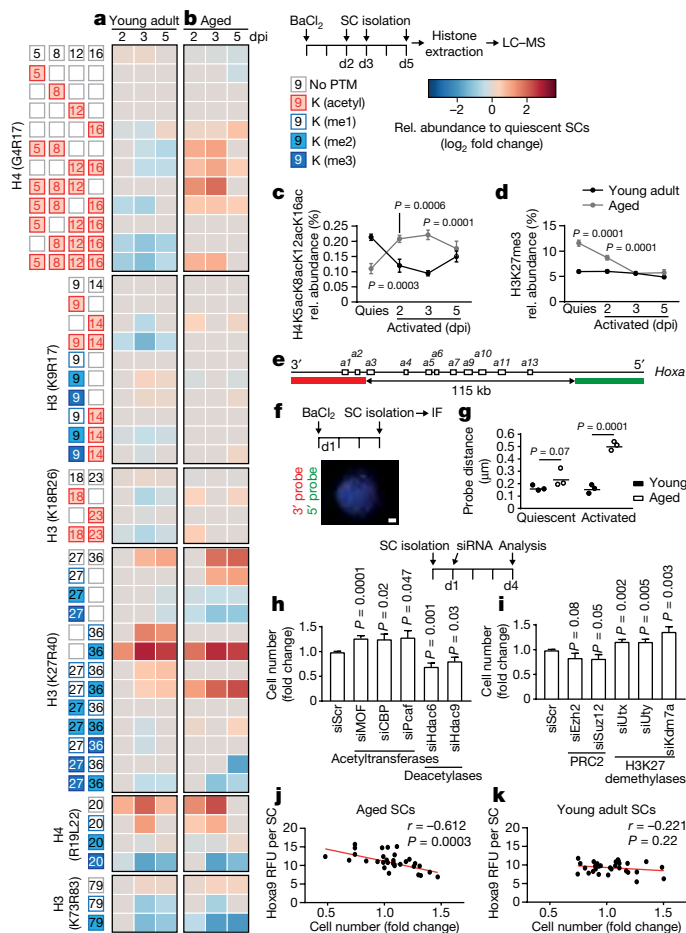


Figure 4 | Altered epigenetic stress response in aged SCs. **a, b,** Heatmap of mass spectrometry (LC-MS) analysis displaying significant ($P < 0.05$) relative changes in abundance of the indicated histone modifications (measured at the indicated peptides) at the indicated days post injury (dpi). **c, d,** Trajectory time-course plots showing relative abundance of H4K5acK8acK12acK16ac (**c**) or H3K27me3 (**d**) in freshly isolated quiescent (quies) or *in vivo* activated SCs purified at indicated time points post muscle injury. **e–g,** Fluorescence *in situ* hybridization of freshly isolated quiescent or *in vivo* activated SCs with the indicated probes spanning the *Hoxa* cluster (**e**); an exemplary image (**f**); and the average probe distance (**g**). Scale bar, 1 μm . **h, i,** Relative changes in SC number 4 days after transfection of freshly isolated SCs from aged mice with the indicated siRNAs. **j, k,** Pearson correlation of relative cell number and *Hoxa9* immunofluorescence signal of SCs from young adult and aged mice 4 days after transfection with a selection of siRNAs targeting different classes of chromatin modifiers. RFU, relative fluorescence units. P values were calculated by two-way ANOVA (**c, d, g**), two-sided Student's *t*-test (**a, b, h, i**), or Pearson correlation (**j, k**). $n = 4$ mice in **a–d**; $n = 3$ mice with 50 nuclei per replicate in **g**; $n = 7$ mice (*Ezh2* siRNA), 8 mice (all others) in **h, i**; $n = 6$ mice (aged), $n = 3$ mice (young) in **j, k**.

H3K27me3; Extended Data Fig. 8a; consistent with ref. 22), and lower amounts of histone modifications typically enriched on active genes (for example, various H4 acetylation motifs, H3K14ac, H3K18ac and H3K36me2; Extended Data Fig. 8a). A time-dependent shift towards a heterochromatic state occurred during SC activation in young adult mice, whereas activation in aged SCs generated the opposite response (Fig. 4a, b). Although selective active marks such as H3 and H4 acetylation motifs declined in SCs from young adult mice during activation, there was a substantial increase in these marks in aged SCs (Fig. 4a–c). Conversely, repressive marks (for example, H3K27me3) decreased in SCs from aged mice but remained stable in SCs from young adult mice during activation (Fig. 4a, b, d). The observed shift of the chromatin towards a more permissive state after SC activation appeared to also

affect the *Hoxa* cluster as this locus displayed an increased chromatin decompaction after SC activation in aged mice but not in young adult mice (Fig. 4e–g).

To analyse the functional contribution of different types of chromatin modifications in activated SCs from aged mice, a set of genetic and pharmacological experiments was conducted. The expression of key enzymes involved in chromatin modifications detected by RNA-sequencing analysis was similar in activated SCs from young adult and aged mice (Extended Data Fig. 8b). However, knockdown of the acetyltransferases *MOF* (also known as *Kat8*), *CBP* (*Crebbp*) or *Pcaf* (*Kat2b*) improved the proliferative capacity of SCs from aged mice in bulk culture, whereas knockdown of histone deacetylases led to a reduction (Fig. 4h). Furthermore, knockdown of the H3K27 demethylases *Utx* (also known as *Kdm6a*), *Uty* or *Kdm7a* promoted the proliferation of aged SCs, which was instead inhibited by knockdown of *Suz12* and *Ezh2* (Fig. 4i), members of the PRC2 protein complex responsible for H3K27me3. Multi-acetylation motifs, as observed in activated SC from aged mice (Fig. 4b, c), are preferred binding sites for bromodomain-containing proteins²³. Eight out of eleven non-toxic bromodomain inhibitors available from the Structural Genomics Consortium exhibited positive effects on the proliferative capacity of SCs from aged mice (Extended Data Fig. 8c, d, $P = 4.2 \times 10^{-4}$). Targeting major classes of chromatin modifiers by a selected set of siRNAs (Supplementary Table 1) revealed a significant inverse correlation ($r = -0.612$) between siRNA-mediated changes in *Hoxa9* protein expression and the proliferative capacity of SCs from aged mice, with no such effects observed in SCs from young adult mice (Fig. 4j, k). Similarly, siRNAs against *MOF* and *Utx* as well as bromodomain inhibitors led to significant decreases in the *Hoxa9* protein level in activated myofibre-associated SCs from aged mice (Extended Data Fig. 8e–g). In summary, activated SCs from aged mice exhibit site-specific and global aberrations in the epigenetic stress response, resulting in *Hoxa9* activation and profound negative effects on SC function, which are ameliorated by targeting the respective enzymes underlying these alterations.

By analysing the downstream effects of *Hoxa9* induction through lentiviral-mediated *Hoxa9* overexpression, we found a strong reduction in the colony forming and proliferative capacity of SCs from young adult mice (Extended Data Fig. 9a–c). The overexpression of other *Hox* genes exerted similar effects (Extended Data Fig. 9d) but the *Hoxa9* results are probably most relevant for physiological ageing because only *Hoxa9* was upregulated in activated SCs from aged mice (Fig. 1). The impaired myogenic capacity of SCs in response to *Hoxa9* overexpression was associated with increased rates of apoptosis and decreased cell proliferation (Extended Data Fig. 9e–h). Furthermore, *Hoxa9* induction associated with the suppression of several cell cycle regulators and induction of cell cycle inhibitors and senescence-inducing genes (Extended Data Fig. 9i) as well as with increased staining for senescence-associated β -galactosidase (Extended Data Fig. 9j, k). Microarray expression analysis of *Hoxa9*-overexpressing SCs compared to controls revealed that among the top 12 pathways regulated by *Hoxa9* were several major developmental pathways that have previously been shown to impair SC function and muscle regeneration in the context of ageing^{2,3,5,6,9,24,25} (Fig. 5a, Extended Data Fig. 9l–o). ChIP analysis of putative *Hoxa9*-binding sites (Supplementary Table 1) in *Hoxa9*-overexpressing primary myoblasts indicated that a high number of these genes are probably direct targets of *Hoxa9* (Extended Data Fig. 9p; cumulative P value over tested genes: $P = 1 \times 10^{-7}$). *Hoxa9* strongly induced downstream targets of the Wnt, TGF β and JAK/STAT pathways, but targeted activation of each one of these pathways alone only led to slight changes in the expression of target genes of the other two pathways (Extended Data Fig. 9q–s), suggesting that *Hoxa9* acts as a central hub required for the parallel induction of these pathways in aged SCs. Of note, the inhibition of *Stat3*, *Bmp4* or *Ctnnb1* (encoding β -catenin) by shRNAs as well as pharmacological inhibition of the Wnt, TGF β or JAK/STAT pathway was sufficient to improve the myogenic colony forming capacity of SCs overexpressing *Hoxa9* (Fig. 5b,

transplantation. Work on this project in K.L.R.'s laboratory was supported by the DGF (RU-745/10, RU-745/12), the ERC (2012-AdG 323136), the state of Thuringia, and intramural funds from the Leibniz association. J.V.M. was supported by a grant from the DFG (MA-3975/2-1). C.F. acknowledges support by the DFG (FE-1544/1-1) and EMBO (long-term postdoctoral fellowship ALTF 55-2015). R.A. was supported by the ERC (AdvGr 670821 (Proteomics 4D)). The funding for the *Hoxa9*^{-/-} mice to K.L.M. was provided by a grant of the NIH (HL096108). R.R. was supported by a grant from the NIH (R01GM106056). This work was further supported by grants to H.A.K. from the DFG (SFB 1074 project Z1), the BMBF (Gerontosys II, Forschungskern SyStaR, project ID 0315894A), and the European Community's Seventh Framework Programme 390 (FP7/2007-2013, grant agreement 602783).

Author Contributions S.S. designed and performed most experiments, analysed data, interpreted results and wrote the manuscript. F.B. designed and performed RNAi, ChIP and FISH experiments on isolated SCs, analysed data, interpreted results and wrote the manuscript. C.F. and R.A. designed and performed LC-MS experiments, analysed data, interpreted results and wrote the manuscript. A.H.B., U.K., H.H., C.S.V. and M.S. performed individual experiments

and analysed data. A.L. performed microarray experiments. D.B.L. provided support and suggestions for ChIP experiments. K.L.M. provided *Hoxa9*^{-/-} mice. J.M.K. and H.A.K. performed microarray and pathway analysis, analysed putative *Hoxa9*-binding sites and provided support for statistical analysis. B.X. and R.R. conducted analysis of putative *Hoxa9*-binding sites. F.N. analysed RNA-sequencing data and performed correlation analysis. J.V.M. and S.T. conceived the project, designed and performed individual experiments, interpreted results and wrote the manuscript. K.L.R. conceived the project, designed experiments, interpreted results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.V.M. (julia.vonmaltzahn@leibniz-fli.de), S.T. (stefan.tuempel@leibniz-fli.de) or K.L.R. (lenhard.rudolph@leibniz-fli.de).

Reviewer Information *Nature* thanks J. Gil and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to estimate sample size. No randomization was used. No animals were excluded. The evaluator was blinded to the identity of the specific sample as much as the nature of the experiment allowed it.

Mice. We purchased female young adult C57/BL6j mice (3–4 months) and aged C57/BL6j mice (22–28 months) from Janvier (wild-type mice). Female and male *Hoxa9*^{−/−} mice have been described²⁸ and were obtained together with age- and gender-matched littermate controls from K. L. Medina. Mice were housed in a pathogen-free environment and fed with a standard diet *ad libitum*. Animal experiments were approved by the Thüringer Landesamt für Verbraucherschutz (Germany) under Reg.-Nr. 03-006/13, 03-012/13 and 03-007/15 and by the Regierungspräsidium Tübingen (Germany) under Reg.-Nr. 35/9185.81-3/919.

Muscle injury. Mice were anaesthetized using isoflurane in air and oxygen through a nose cone. For SC activation, muscles were injured by injecting a total volume of 50 µl of 1.2% BaCl₂ (Sigma) into approximately 20 sites in the hindlimb muscles. For regeneration and transplantation experiments, tibialis anterior muscle of the right leg was injected with 50 µl cardiotoxin (CTX, 10 µM, Sigma).

SC isolation and FACS. Muscles from hindlimbs from young adult or aged mice were dissected and collected in PBS on ice. Muscles were rinsed with PBS, minced with scissors and incubated in DMEM with Collagenase (0.2%, Biochrom) for 90 min at 37 °C and 70 r.p.m. Digested muscles were washed with 10% FBS in PBS, triturated and incubated in Collagenase (0.0125%) and Dispase (0.4%, Life Technologies) for 30 min at 37 °C and 100 r.p.m. The muscle slurry was diluted with 10% FBS in PBS, filtered through 100-µm cell strainers and spun down at 500g for 5 min. Cell pellets were resuspended in FACS buffer (2% FBS in HBSS) and filtered through 40-µm cell strainers and pelleted at 500g for 5 min. Pellets were resuspended in FACS buffer and stained with anti-mouse CD45 PE conjugate (30-F11, eBioscience), anti-mouse CD11b PE conjugate (M1/70, eBioscience), anti-mouse Sca-1 PE conjugate (D7, BioLegend), anti-mouse CD31 PE/Cy7 conjugate (390, BioLegend) and anti-mouse α7-integrin Alexa Fluor 647 conjugate (R2F2, AbLab) for 20 min at 4 °C on a rotating wheel. Cells were washed with FACS buffer. Live cells were identified as calcein blue positive (1:1,000, Invitrogen) and propidium iodide negative (PI, 1 µg ml^{−1}, BD Biosciences). SCs were identified as CD45[−]Sca-1⁺CD11b[−]CD31[−]α7-integrin⁺. Cell sorting was performed on a FACS Aria III with Diva Software (BD).

Culture of SCs. SCs and SC-derived primary myoblasts were cultured at 37 °C, 5% CO₂, 3% O₂ and 95% humidity in growth medium on collagen/laminin-coated tissue culture plates for the indicated time periods. Growth medium was comprised of F10 (Life Technologies) with 20% horse serum (GE), 1% penicillin/streptomycin (Life Technologies) and 5 ng ml^{−1} bFGF (Sigma). For coating, tissue culture plates were incubated with 1 mg ml^{−1} collagen (Sigma) and 10 mg ml^{−1} laminin (Life Technologies) in ddH₂O for at least 1 h at 37 °C and allowed to air-dry. For passaging or FACS analysis, cultured cells were incubated with 0.5% trypsin in PBS for 3 min at 37 °C and collected in FACS buffer. Treatment of SCs with noggin (Peprotech) or DKK1 (Peprotech) was done at 100 ng ml^{−1} concentration. SCs and SC-derived primary myoblasts were treated with 1 µM of chemical probes provided by the Structural Genomics Consortium (SGC, <http://www.thesgc.org/>; chemical-probes/epigenetics)^{29,30}. OICR-9429 and bromodomain inhibitors were described previously^{19,31–39}.

Clonal myogenesis assay. Freshly isolated SCs from young adult and aged mice were sorted in growth medium in 96-well plates using the automated cell deposition unit of the FACS Aria III. After 5 days, wells containing myogenic colonies were counted by brightfield microscopy. For clonal analysis of lentivirus-transduced SCs, infected (eGFP⁺ and/or BFP⁺) live (DAPI[−]) cells were sorted as one cell per well in growth medium and wells containing myogenic colonies were counted by fluorescence microscopy (Axio Observer, Zeiss) after 5 days. A colony was defined by the presence of at least two cells.

Alamar blue assay. SCs or SC-derived primary myoblasts were seeded at 500 cells per well in growth medium into 96-well plates. After 4 days of culture, the viability was measured by adding Alamar Blue (Life Technologies) as 10% of the sample volume. Cells were incubated for 2 h at 37 °C and fluorescence intensity was measured at an excitation/emission wavelength of 560/590 nm.

BrdU assay. SCs were incubated with 5 µM BrdU (Sigma) in growth medium for 2 h. Cells were fixed with 4% PFA, permeabilized with 0.5% Triton X-100 and incubated with 2 N HCl/PBS for 30 min at room temperature. Incorporated BrdU was detected using anti-BrdU (347580, BD Biosciences) and Alexa-594 fluorochrome (Life Technologies) for 1 h at room temperature. Nuclei were counterstained with DAPI in PBS.

TUNEL assay. TUNEL assay was performed using the *In situ* Cell Death Detection Kit, Texas Red (Roche) according to the manufacturer's instructions.

Senescence-associated β-galactosidase assay. SCs were fixed in 4% PFA and stained with staining solution (5 mM potassium ferricyanide, 5 mM potassium

ferricyanide, 2 mM MgCl₂, 150 mM NaCl, 1 mg ml^{−1} X-Gal) in citrate/sodium-phosphate buffer (pH 6) overnight at 37 °C. Staining solution was removed by rinsing several times with PBS.

Myofibre isolation and culture. Individual myofibres were isolated from the extensor digitorum longus muscle as described previously^{40,41}. Isolated myofibres were cultured in DMEM containing 20% FBS and 1% chicken embryo extract (Biomol) in dishes coated with horse serum. Freshly isolated fibres or fibres cultured for 24–34 h and 72 h were fixed with 2% PFA and subjected to immunofluorescence analysis. Clusters of SCs were counted on at least 10–15 fibres per replicate. A cluster was defined by the presence of at least three adjacent cells. For quantification of immunofluorescence staining of myofibre-associated quiescent and activated SCs, at least 20 fibres were analysed per replicate. Treatment of myofibre-associated SCs with chemical probes provided by the Structural Genomics Consortium (SGC) was done 4 h after isolation at 1 µM concentration.

siRNA transfection. Transfection of SCs was performed in a reverse manner: SCs were seeded in growth medium into individual wells of a 384-well plate pre-filled with transfection mix. For floating cultures of single myofibres, transfections were performed 4 h after isolation in myofibre culture medium. Transfections were done using Lipofectamin RNAiMAX (Life Technologies) according to manufacturer's instructions. For gene knockdown either Silencer Select siRNAs (Life Technologies) or ON-TARGETplus siRNA SMART-pools (Dharmacon) were used. Respective Silencer Select or ON-TARGETplus SMART-pool non-targeting siRNAs were used as negative control. siRNA sequences are listed in Supplementary Table 1. Transfection efficiency was monitored using a Cy3-labelled control siRNA (Life Technologies). After transfection, FACS-sorted SCs or myofibre-associated SCs were cultured for the indicated time periods and fixed in 2% PFA in PBS. *In vivo* knockdown experiments were performed as described earlier⁴¹. siRNA sequences were modified to the Accell self-delivering format (Dharmacon) and 100 µg Accell siRNA were injected into tibialis anterior muscle 2 days after CTX injury. *In vivo* knockdown was evaluated from SCs isolated from injected tibialis anterior muscle 3 days after transfection. Transfected muscles were collected 5 days after siRNA injection, frozen in 10% sucrose/OCT in liquid nitrogen and stored at −80 °C.

Lentivirus production and transduction. Lentivirus was produced in Lenti-X cells (Clontech) after co-transfection of 15 µg shRNA or cDNA plasmid, 10 µg psPAX2 helper plasmid and 5 µg pMD2.G according to standard procedures⁴². Virus was concentrated by centrifugation for 2.5 h at 106,800g and 4 °C, and virus pellet was resuspended in sterile PBS. Lentiviral transduction was carried out in growth medium supplemented with 8 µg ml^{−1} polybrene (Sigma).

Plasmids. cDNA was inserted into the SF-LV-cDNA-eGFP plasmid⁴³. Primers used for cloning of individual *Hox* cDNAs are listed in Supplementary Table 1. shRNA was inserted into the SF-LV-shRNA-eGFP plasmid using mir30 primers (Supplementary Table 1). shRNA sequences are listed in Supplementary Table 1.

SC transplantation. SCs were FACS purified and transduced with a lentivirus on Retrofectin (Takara) coated 48-well plates⁴. After 8–10 h, SCs were obtained by resuspension and washed several times with FACS buffer. For each engraftment, 10,000 SCs were resuspended in 0.9% NaCl and immediately transplanted into tibialis anterior muscles of adult immunosuppressed mice that had been injured with CTX 2 days before. Immunosuppression with FK506 (5 mg kg^{−1} body weight, Sigma) was started at the day of injury using osmotic pumps (model 2004, Alzet) and maintained throughout the entire time of engraftment. Engrafted muscles were collected 3 weeks after transplantation and fixed in 4% PFA for 30 min at room temperature followed by incubation in 30% sucrose/PBS overnight at 4 °C. Fixed muscles were frozen in 10% sucrose/OCT in liquid nitrogen and stored at −80 °C.

Immunohistochemistry. Cryosections of 10 µm were cut from frozen muscle using the Microm HM 550. Cryosections were rinsed once with PBS and fixed in 2% PFA in PBS for 5 min at room temperature. Sections were rinsed three times for 5 min with PBS, permeabilized with 0.5% Triton X-100/0.1 M glycine in PBS for 5 min at room temperature followed again by rinsing them three times with PBS. Sections were blocked in PBS supplemented with 5% horse serum and 1:40 mouse on mouse blocking reagent (Vector labs) for 1 h at room temperature. Incubation with primary antibodies was carried out overnight at 4 °C. The next day, sections were rinsed three times with PBS followed by incubation with secondary antibodies for 1 h at room temperature. Sections were rinsed again with PBS and nuclei were counterstained with 1:1,000 DAPI in PBS before mounting with Permafluor (Thermo Scientific). Slides were stored at 4 °C until analysis. The following primary antibodies were used: 1:1,000 chicken anti-GFP (ab6556, AbCam), 1:1,000 rabbit anti-laminin (L9393, Sigma), 1:200 rabbit anti-Ki67 (ab15580, AbCam), undiluted mouse anti-Pax7 (DSHB). The following secondary antibodies were used at 1:1,000: anti-chicken IgG Alexa-Fluor 488, anti-rabbit IgG Alexa-Fluor 488, anti-mouse IgG1 Alexa-Fluor 594 (Life Technologies).

Immunofluorescence. Freshly isolated SCs were allowed to settle on poly-L-lysine-coated diagnostic microscope slides for 30 min at room temperature. All cells and myofibres were fixed with 2% PFA, permeabilized with 0.5% Triton X-100 and blocked with 10% horse serum in PBS for 1 h at room temperature. Cells and fibres were stained with primary antibodies in blocking solution overnight at 4 °C. Samples were washed three times with PBS and incubated with secondary antibodies for 1 h at room temperature. Nuclei were counterstained with DAPI. Cultured cells were kept in PBS; freshly isolated SCs and myofibres were mounted with Permafluor. The following primary antibodies were used: undiluted mouse anti-Pax7 (DSHB), 1:300 rabbit anti-Hoxa9 (07-178, Millipore), 1:500 mouse anti-Mll1 (05-765, Millipore), 1:500 rabbit anti-Wdr5 (A302-429A, Bethyl Laboratories), 1:300 rabbit anti-H3K4me3 (C15410003-50, Diagenode), 1:200 rabbit anti-MyoD (sc-304, Santa Cruz). The following secondary antibodies were used at 1:1,000: anti-rabbit IgG Alexa-Fluor 488, anti-mouse IgG Alexa-Fluor 594, anti-mouse IgG1 Alexa-Fluor 594 (Life Technologies).

Fluorescence in situ hybridization (FISH). Chromatin compaction FISH was done as described previously⁴⁴. DNA of the 3' and 5' probe (Fosmid clones WIBR1-1312N03 and WIBR1-2209G09, CHORI) was labelled with digoxigenin or biotin by nick-translation (Roche). 100 ng of probe DNA was used per slide, together with 5 µg mouse CotI DNA (Life Technologies) and 5 µg single-stranded DNA (Ambion). Approximately 5,000 freshly sorted SCs were allowed to settle on poly-L-lysine-coated diagnostic microscope slides for 30 min at room temperature and were fixed with 2% PFA for 5 min. After washing three times with PBS, slides were incubated with 0.1 M HCl for 5 min and permeabilized with 0.5% Triton X-100 in 0.5% saponin for 10 min before freeze–thaw in 20% glycerol in PBS. Denaturation was performed in 50% formamide, 1% Tween-20 and 10% dextran sulfate/2× SSC for 5 min at 75 °C before applying the hybridization cocktail. Probes were hybridized overnight at 37 °C in a humidified chamber. Slides were rinsed three times with 2× SSC, blocked with 2% BSA in 0.1% Tween-20 in PBS for 1 h at room temperature, and hybridized probes were visualized with anti-digoxigenin-rhodamine (S7165, Millipore) and Streptavidin-Cy2 (016-220-084, IR USA) for 30 min at room temperature. Nuclei were counterstained with DAPI.

Digital image acquisition and processing. Immunofluorescence images of muscle sections, myofibres and freshly isolated SCs were acquired using the upright microscope Axio Imager (Zeiss) with 10×, 20× and 100× objectives and a monochrome camera. Brightfield and immunofluorescence images of cultured SCs were captured using the microscope Axio Observer (Zeiss) with 5×, 10× and 20× objectives and a monochrome camera. Image acquisition and processing was performed using the ZEN 2012 software (Zeiss). Brightness and contrast adjustments were applied to the entire image before the region of interest was selected. For the analysis of muscle sections, several images covering the whole area of the section were acquired in a rasterized manner and assembled in Photoshop CS6 (Adobe) to obtain an image of the entire section. Images were analysed using ImageJ software. The number of Pax7⁺ cells in regeneration experiments was normalized to the area of the entire muscle section. CTCF was determined for each SC using the calculation: integrated density – (area of selected cell × mean fluorescence of background readings) (ref. 45).

RNA isolation and reverse transcription. Total RNA was isolated from freshly FACS-isolated or cultured SCs by using the MagMAX 96 total RNA Isolation Kit (Ambion) according to the manufacturer's protocol. The GoScript Reverse Transcription System (Promega) was used for cDNA synthesis from total RNA according to manufacturer's instructions.

ChIP. 5 × 10⁴–1 × 10⁵ cells were crosslinked in 1% formaldehyde (Thermo Scientific) for 10 min. Crosslinking was quenched with glycine and cells were washed two times with ice-cold PBS. For ChIP of H3K4me3, cells were lysed in lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8.1, 1× Roche cComplete Protease Inhibitor) and chromatin was sonicated in Snap Cap microTUBEs using a Covaris M220 sonicator to a fragment size of 150–300 bp. Chromatin was cleared for 10 min at 17,000g, and one-tenth of the chromatin was removed as input fraction. Chromatin was immunoprecipitated overnight with 20 µl Protein A/G bead mix (1:1, Dynabeads, Invitrogen) pre-coupled with 1 µg antibody (C15410003-50, Diagenode) in ChIP-dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, 16.7 mM Tris-HCl pH 8.1, 1× Roche cComplete Protease Inhibitor). Beads were washed three times with low-salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl, Tris-HCl, pH 8.1) and three times with LiCl buffer (350 mM LiCl, 1% IPEGAL CA630, 1% deoxycholic acid, 1 mM EDTA, 10 mM Tris-HCl, pH 8.1). For ChIP of Mll1, Wdr5 or haemagglutinin (HA)-tagged Hoxa9 cells were resuspended in sonication buffer (0.1% SDS, 1% Triton X-100, 0.1% Na-deoxycholate, 1 mM EDTA, 140 mM NaCl, 50 mM HEPES, pH 7.9), incubated on ice for 10 min and sonicated to a fragment size of 300–600 bp as described above. Chromatin was cleared for 10 min at 17,000g and unspecific binding was absorbed with 5 µl of Protein G beads for 1 h. One-tenth (Mll1/Wdr5) or one-twentieth

(HA-tag) of the chromatin was removed as input fraction. Chromatin was immunoprecipitated overnight with 2 µg of antibody (Mll1: A300-086A, Wdr5: A302-429A, Bethyl Laboratories; HA-tag: ab9110, Abcam). Chromatin-antibody complexes were captured with 20 µl Protein A/G bead mix (1:1, Dynabeads, Invitrogen) for 2 h. Beads were washed twice with sonication buffer, twice with NaCl buffer (0.1% SDS, 1% Triton X-100, 0.1% Na-deoxycholate, 1 mM EDTA, 500 mM NaCl, 50 mM HEPES, pH 7.9), twice with LiCl buffer and once with TE buffer. Decrosslinking and elution was performed in 50 µl decrosslinking buffer (1% SDS, 100 mM NaHCO₃, 250 mM NaCl) for 4 h at 65 °C with continuous shaking and subsequent Proteinase K treatment for 1 h at 45 °C. DNA was purified using Agencourt AMPure XP beads (Beckman Coulter) with a beads:sample ratio of 1.8:1 or MinElute PCR Purification Kit according to manufacturer's protocols. **Quantitative PCR.** Quantitative PCR (qPCR) was performed with an ABI 7500 Real-Time PCR System (Applied Biosystems) in technical duplicates from the indicated number of biological replicates. The qPCR was carried out in a volume of 12 µl using the Absolute qPCR Rox Mix (Thermo Scientific) and the Universal Probe Library (Roche). Primer and probe sets for the detection of single genes are listed in Supplementary Table 1. *Gapdh* was detected with rodent *Gapdh* control reagents (Applied Biosystems). Relative expression values were calculated using the $\Delta\Delta C_t$ method.

$$\Delta C_t = C_t[\text{gene of interest}] - C_t[\text{Gapdh}]$$

$$\text{Relative expression} = 2^{(-\Delta C_t)}$$

qPCR analysis of ChIP samples was performed using SYBR Green Supermix (Biorad) in a final reaction volume of 10 µl and 0.75 µM final primer concentration. Primers are listed in Supplementary Table 1. HA-tag ChIP signals were calculated as percentage of the input fraction. The $\Delta\Delta C_t$ method was used to calculate fold enrichment of a genomic locus over the ChIP specific background control (*Actb* intergenic region for H3K4me3 or gene desert for *Mll1* and *Wdr5*), both normalized to the signal in the input fraction:

$$\Delta C_t[\text{normalized to input}] = (C_t[\text{ChIP}] - (C_t[\text{input}] - \log_2(\text{input dilution factor})))$$

$$\Delta\Delta C_t = \Delta C_t[\text{region of choice normalized to input}] - \Delta C_t[\text{control region normalized to input}]$$

$$\text{Fold enrichment} = 2^{(-\Delta\Delta C_t)}$$

Nanostring analysis. Pellets of freshly isolated SCs were lysed with 3 µl RLT buffer (QIAGEN) and subjected to Nanostring analysis according to manufacturer's instructions using a custom-made Hox gene nCounter Elements TagSet (Nanostring Technologies). Relative expression to the housekeeping genes *Gapdh*, *Hmbs* and *Polr2a* was calculated using nSolver Software (v2.0) after background correction and normalization to hybridized probe signals.

Proteomic analysis of histone modifications. Preparation of histones for mass spectrometry, data acquisition and analysis were essentially performed as described previously²¹ with modifications described below. In brief, histones were isolated by acid extraction, derivatised by d6-acetic anhydride (CD₃CO, Aldrich) and digested with sequencing-grade trypsin (Promega) overnight at a trypsin:protein ratio of 1:20. To acetylate free peptide N termini, trypsinised histones were derivatised again for 45 min at 37 °C using 1:20 (v/v) d6-acetic anhydride (CD₃CO, Aldrich) in 50 mM ammonium bicarbonate buffered to pH 8 by ammonium hydroxide solution. After derivatization, peptides were evaporated in a speed-vac at 37 °C to near dryness, resuspended in 50 µl of 0.1 formic acid and purified by a StageTip protocol using two discs of C18 followed by one disc of activated carbon (3 M Empore). After StageTip purification, the samples were evaporated in a speed-vac to near dryness, resuspended in 20 µl of 0.1% formic acid and stored at –20 °C until mass spectrometry acquisition. The histone samples were separated on a reversed-phase liquid chromatography column (75-µm, New Objective) that was packed in-house with a 15-cm stationary phase (ReproSil-Pur C18-AQ, 1.9 µm). The column was connected to a nano-flow HPLC (EASY-nLC 1000; Thermo Scientific) and peptides were electrosprayed in a Q Exactive mass spectrometer (Thermo Fisher Scientific). Buffer A was composed of 0.1% formic acid in HPLC-grade water and buffer B was 0.1% formic acid in ACN. Peptides were eluted in a linear gradient with a flow rate of 300 nl per minute, starting at 3% B and ramping to 35% in 52 min, followed by an increase to 50% B in 4 min, followed by an increase to 98% in 4 min and then holding at 98% B for another 6 min. Mass spectrometry was operated in a combined shotgun-PRM mode targeting positional isomers. Ion chromatograms were extracted with Thermo Xcalibur and Skyline and data summarization and statistical analysis was performed in Excel and R. Relative abundances were calculated from the raw signal reads, according to the formulas described previously²¹ without further normalizations.

Microarray and bioinformatics analysis. Gene expression analysis was performed using the Mouse GE 8x60K Microarray Kit (Agilent Technologies, Design ID 028005). 100 ng total RNA isolated from SCs were used for the labelling. Samples were labelled with the Low Input Quick Amp Labelling Kit (Agilent Technologies) according to the manufacturer's instructions. Slides were scanned using a microarray scanner (Agilent Technologies). Expression data were extracted using the Feature Extraction software (Agilent Technologies). Preprocessing of expression data was performed according to Agilent's standard workflow. Using five quality flags (gIsPosAndSignif, gIsFeatNonUnifOL, gIsWellAboveBG, gIsSaturated, and gIsFeatPopnOL) from the Feature Extraction software output, probes were labelled as detected, not detected, or compromised. Gene expression levels were background corrected, and signals for duplicated probes were summarized by geometric mean of non-compromised probes. After \log_2 transformation, a percentile shift normalization at the 75% level and a baseline shift to the median baseline of all probes was performed. All computations were performed using the R statistical software framework (<http://www.R-project.org>). Differentially expressed genes were calculated by the shrinkage T-statistic⁴⁶ and controlled for multiple testing by maintaining a false discovery rate (FDR) < 0.05 (ref. 47).

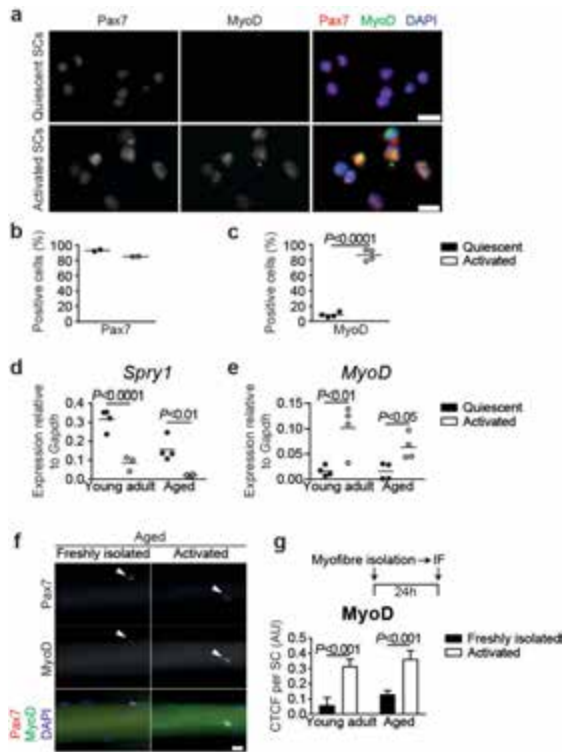
RNA-sequencing analysis. Sequencing reads were filtered out for low quality sequences and trimmed of low quality bases by using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Mapping to mm9 genome was performed by using TopHat software⁴⁸. Gene quantification was performed by using HT-Seq and differentially expressed genes (DEGs) were estimated by using DESEQ2 (refs 49, 50) within the R statistical software framework (<http://www.R-project.org>) with $P < 0.01$. Pearson correlation heatmaps were generated by using custom R scripts by selecting genes having more than 10 read counts in all the samples of at least one condition and an interquartile range (IQR) > 0.5. Significance of overlapping DEGs was calculated by normal approximation of hypergeometric probability.

Identification of Hoxa9-binding sites. Transcription start and end sites of putative Hoxa9 target genes were collected from the UCSC Genome Browser⁵¹ with mm8 track. Sequences in gene body regions (from transcription start to end sites), promoter regions (−2/+1 kb relative to transcription start sites), and distal intergenic regions (−50/+50 kb relative to transcription start sites) of 26 genes were prepared for identification of Hoxa9 binding sites. These sequences were aligned based on the previously reported consensus motifs for Hoxa9-Meis1-Pbx1 (ATGATTTATGGC)⁵² and Meis1 (TGTC)⁵³. Putative Hoxa9-binding sites were aligned when they contained either no mismatch or one mismatch, and Meis1 motifs were aligned with no mismatch allowed. Hoxa9-binding sites with at least one Meis1-binding site within 300 bp on the same DNA strand were selected for further analysis. Identified Hoxa9-binding sites are listed in Supplementary Table 1.

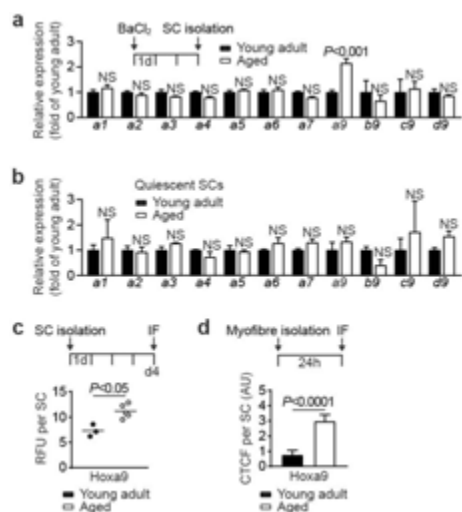
Statistics. If not stated otherwise, results are presented as mean and s.e.m. from the number of samples indicated in the figure legends. Two groups were compared by two-sided Student's *t*-test or two-sided Mann–Whitney *U*-test. For multiple comparisons a two-way ANOVA was performed using a FDR < 0.5 to correct for multiple comparisons. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$. Statistical analysis was done using GraphPad Prism 6 software and R (v3.3.1).

Data availability statement. Microarray and RNA-sequencing data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE87812. Further data that support the findings of this study are available from the corresponding authors upon reasonable request. Source data for the Figures and Extended Data Figures are provided with the paper.

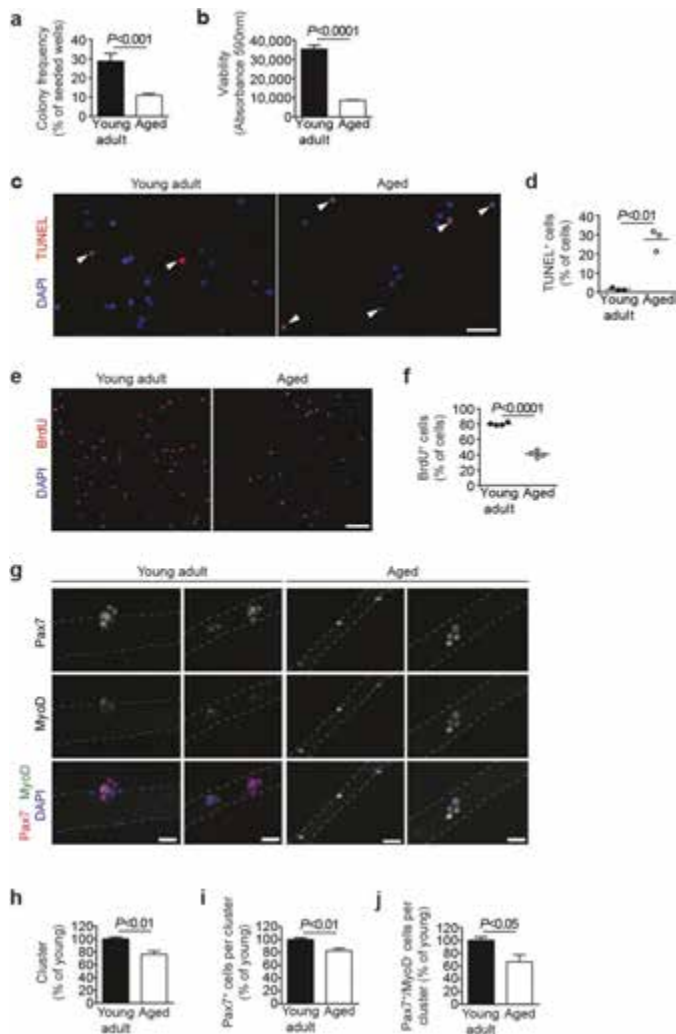
28. Lawrence, H. J. *et al.* Mice bearing a targeted interruption of the homeobox gene *HoxA9* have defects in myeloid, erythroid, and lymphoid hematopoiesis. *Blood* **89**, 1922–1930 (1997).
29. Brown, P. J. & Müller, S. Open access chemical probes for epigenetic targets. *Future Med. Chem.* **7**, 1901–1917 (2015).
30. Barsyte-Lovejoy, D. *et al.* Chemical biology approaches for characterization of epigenetic regulators. *Methods Enzymol.* **574**, 79–103 (2016).
31. Theodoulou, N. H. *et al.* Discovery of I-BRD9, a selective cell active chemical probe for bromodomain containing protein 9 inhibition. *J. Med. Chem.* **59**, 1425–1439 (2016).
32. Picaud, S. *et al.* Generation of a selective small molecule inhibitor of the CBP/p300 bromodomain for leukemia therapy. *Cancer Res.* **75**, 5106–5119 (2015).
33. Picaud, S. *et al.* PFI-1, a highly selective protein interaction inhibitor, targeting BET bromodomains. *Cancer Res.* **73**, 3336–3346 (2013).
34. Martin, L. J. *et al.* Structure-based design of an *in vivo* active selective BRD9 inhibitor. *J. Med. Chem.* **59**, 4462–4475 (2016).
35. Hay, D. A. *et al.* Discovery and optimization of small-molecule ligands for the CBP/p300 bromodomains. *J. Am. Chem. Soc.* **136**, 9308–9319 (2014).
36. Drouin, L. *et al.* Structure enabled design of BAZ2-ICR, a chemical probe targeting the bromodomains of BAZ2A and BAZ2B. *J. Med. Chem.* **58**, 2553–2559 (2015).
37. Clark, P. G. *et al.* LP99: discovery and synthesis of the first selective BRD7/9 bromodomain inhibitor. *Angew. Chem.* **127**, 6315–6319 (2015).
38. Chen, P. *et al.* Discovery and characterization of GSK2801, a selective chemical probe for the bromodomains BAZ2A and BAZ2B. *J. Med. Chem.* **59**, 1410–1424 (2016).
39. Filippakopoulos, P. *et al.* Selective inhibition of BET bromodomains. *Nature* **468**, 1067–1073 (2010).
40. Pasut, A., Jones, A. E. & Rudnicki, M. A. Isolation and culture of individual myofibers and their satellite cells from adult skeletal muscle. *J. Vis. Exp.* **73**, 50074 (2013).
41. Bentzinger, C. F. *et al.* Fibronectin regulates Wnt7a signaling and satellite cell expansion. *Cell Stem Cell* **12**, 75–87 (2013).
42. Schambach, A. *et al.* Lentiviral vectors pseudotyped with murine ecotropic envelope: increased biosafety and convenience in preclinical research. *Exp. Hematol.* **34**, 588–592 (2006).
43. Wang, J. *et al.* A differentiation checkpoint limits hematopoietic stem cell self-renewal in response to DNA damage. *Cell* **148**, 1001–1014 (2012).
44. Chambeyron, S. & Bickmore, W. A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* **18**, 1119–1130 (2004).
45. Burgess, A. *et al.* Loss of human Greatwall results in G2 arrest and multiple mitotic defects due to deregulation of the cyclin B-Cdc2/PP2A balance. *Proc. Natl Acad. Sci. USA* **107**, 12564–12569 (2010).
46. Opgen-Rhein, R. & Strimmer, K. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.* **6**, <http://dx.doi.org/10.2202/1544-6115.1252> (2007).
47. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 12 (1995).
48. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
49. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
51. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
52. Shen, W. F. *et al.* HOXA9 forms triple complexes with PBX2 and MEIS1 in myeloid cells. *Mol. Cell. Biol.* **19**, 3051–3061 (1999).
53. Huang, Y. *et al.* Identification and characterization of Hoxa9 binding sites in hematopoietic cells. *Blood* **119**, 388–398 (2012).



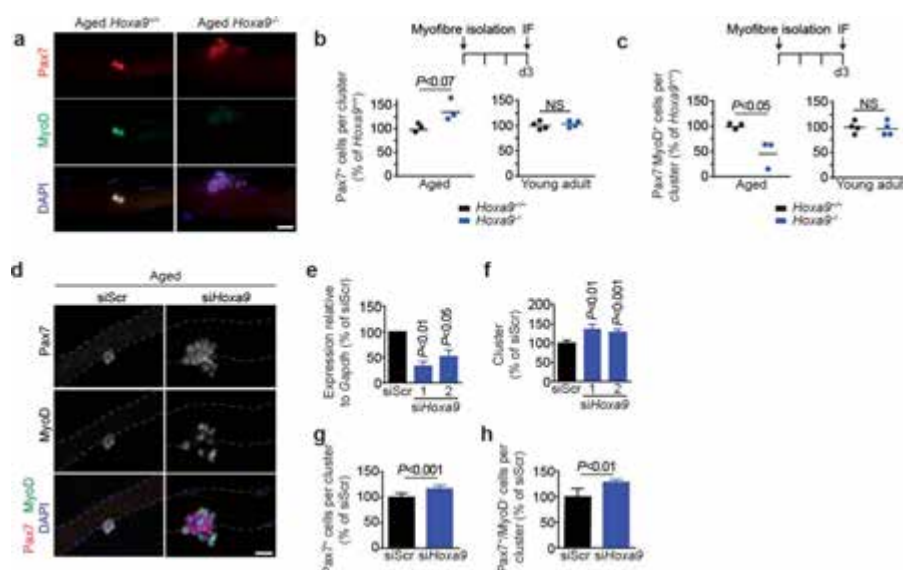
Extended Data Figure 1 | SC activation. **a**, Immunofluorescence staining for Pax7 and MyoD of freshly isolated SCs from injured (activated SCs) and uninjured muscles (quiescent SCs) from young adult mice. Nuclei were counterstained with DAPI (blue). **b**, **c**, Quantification of Pax7⁺ cells (**b**) and MyoD⁺ cells (**c**) in **a**. **d**, **e**, qPCR analysis of *Spry1* (**d**) and *MyoD* (**e**) expression in freshly isolated quiescent and *in vivo* activated SCs of young adult and aged mice. **f**, Immunofluorescence staining for Pax7 and MyoD on freshly isolated and 24-h cultured myofibre-associated SCs from aged mice. Nuclei were counterstained with DAPI (blue). **g**, Corrected total cell fluorescence (CTCF) for MyoD per SC as in **f**. Scale bars, 10 μ m (**a**) and 20 μ m (**f**). *P* values were calculated by two-sided Student's *t*-test (**b**, **c**) or two-way ANOVA (**d**, **e**, **g**). *n* = 2 mice in **b**; *n* = 4 mice in **c**; *n* = 3 mice (young activated), *n* = 4 mice (all others) in **d**; *n* = 4 mice in **e**; *n* = 33/24 nuclei (young), *n* = 35/20 nuclei (aged) from 3 mice in **g**.



Extended Data Figure 2 | Expression of Hox genes in SCs. **a, b**, Nanostring analysis of mRNA expression of *Hoxa* genes and *Hoxa9* paralogues (*b9-c9-d9*) in *in vivo* activated (**a**) and quiescent (**b**) freshly isolated SCs from young adult and aged mice. **c**, Relative fluorescence units (RFU) for *Hoxa9* per SC in 4-day cultured SCs from young adult and aged mice. **d**, Corrected total cell fluorescence (CTFC) for *Hoxa9* per activated SC on 24-h cultured myofibres as in Fig. 1d. *P* values were calculated by two-way ANOVA (**a, b**) or two-sided Mann-Whitney *U*-test (**c, d**). *n* = 3 mice in **a, b**; *n* = 3 mice (young), *n* = 5 mice (aged) in **c**; *n* = 34 nuclei (young), *n* = 32 nuclei (aged) from 4 mice in **d**.

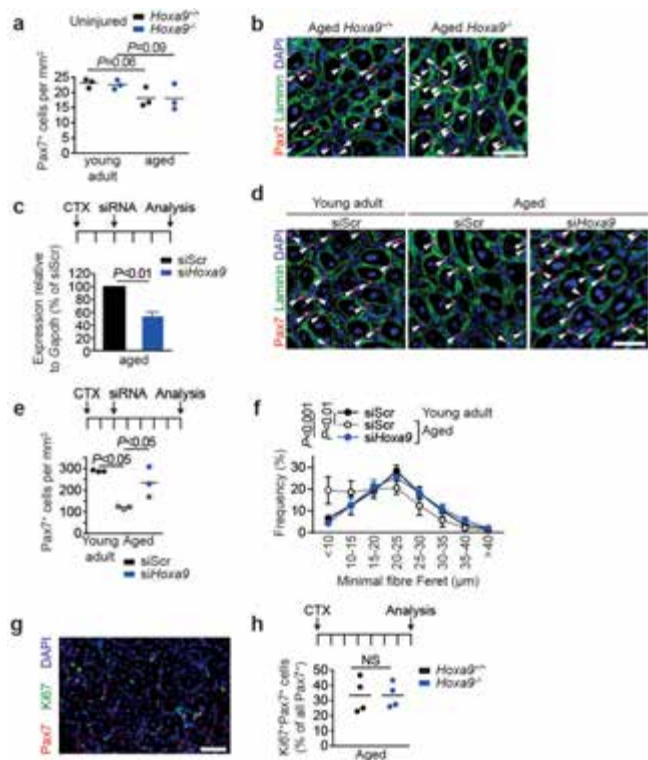


Extended Data Figure 3 | Functional decline in aged SCs. **a**, SCs from young adult and aged mice were sorted as single cells. After 5 days, the frequency of myogenic colonies was assessed. **b**, Equal numbers of FACS-isolated SCs from young adult and aged mice were cultured for 4 days and Alamar Blue assay was performed. **c**, TUNEL staining of SCs isolated from young adult or aged mice after 4 days of culture. Nuclei were counterstained with DAPI (blue). **d**, Quantification of apoptosis based on TUNEL staining in **c**. **e**, BrdU staining of SCs isolated from young adult or aged mice after 4 days of culture. Nuclei were counterstained with DAPI (blue). **f**, Quantification of proliferation based on BrdU staining in **e**. **g**, Immunofluorescence staining for Pax7 and MyoD on myofibres isolated from young adult and aged mice after 72 h in culture. Nuclei were counterstained with DAPI (blue). **h–j**, Quantification of the number of SC-derived clusters with at least 3 adjacent cells (**h**), average number of all Pax7⁺ cells (**i**), or proportion of Pax7⁺/MyoD⁻ cells (**j**) within clusters as in **g**. Scale bars, 20 μ m (**c, g**) and 50 μ m (**e**). *P* values were calculated by two-sided Student's *t*-test. $n = 8$ mice (young), $n = 10$ mice (aged) in **a**; $n = 7$ mice (young), $n = 5$ mice (aged) in **b**; $n = 3$ mice in **d**; $n = 4$ mice in **f**; $n = 4$ mice (aged) in **j**, $n = 5$ mice (all others) in **h–j**.

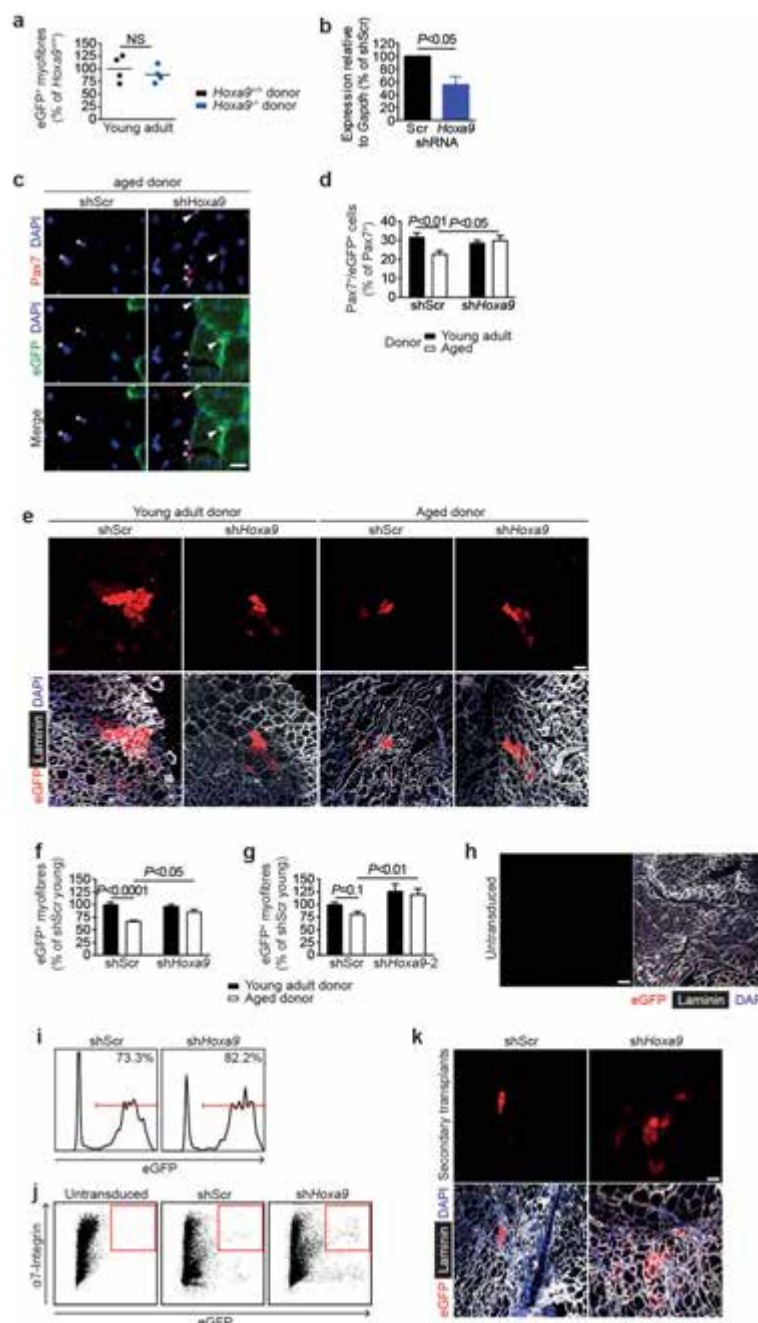


Extended Data Figure 4 | Deletion or knockdown of *Hoxa9* improves SC function in myofibre cultures. **a**, Immunofluorescence staining for Pax7 and MyoD on 72 h cultured myofibre-associated SCs from aged *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice. **b**, **c**, Average number of all Pax7⁺ cells (**b**) or Pax7⁺/MyoD⁺ cells (**c**) within clusters from aged or young adult *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice as shown in **a**. **d**, Immunofluorescence staining for Pax7 and MyoD on 72-h cultured myofibres isolated from aged mice transfected with *Hoxa9* or scrambled (Scr) siRNAs. Nuclei were counterstained with DAPI (blue). **e**, qPCR analysis of *Hoxa9* expression

in SCs transfected with *Hoxa9* siRNA or scrambled control. Two *Hoxa9* siRNAs with different target sequences (Supplementary Table 1) were used. **f–h**, Analysis of 72-h cultured myofibre-associated SCs from **d**. Quantification of the number of SC-derived clusters with at least 3 adjacent cells (**f**), average number of all Pax7⁺ cells (**g**), or proportion of Pax7⁺/MyoD⁺ cells (**h**) within clusters. Scale bars, 20 μ m (**a**, **d**). Dashed lines outline myofibres. *P* values were calculated by two-sided Student's *t*-test. *n* = 3 mice (aged), *n* = 4 mice (young) in **b**, **c**; *n* = 3 mice in **e**; *n* = 5 mice in **f–h**.

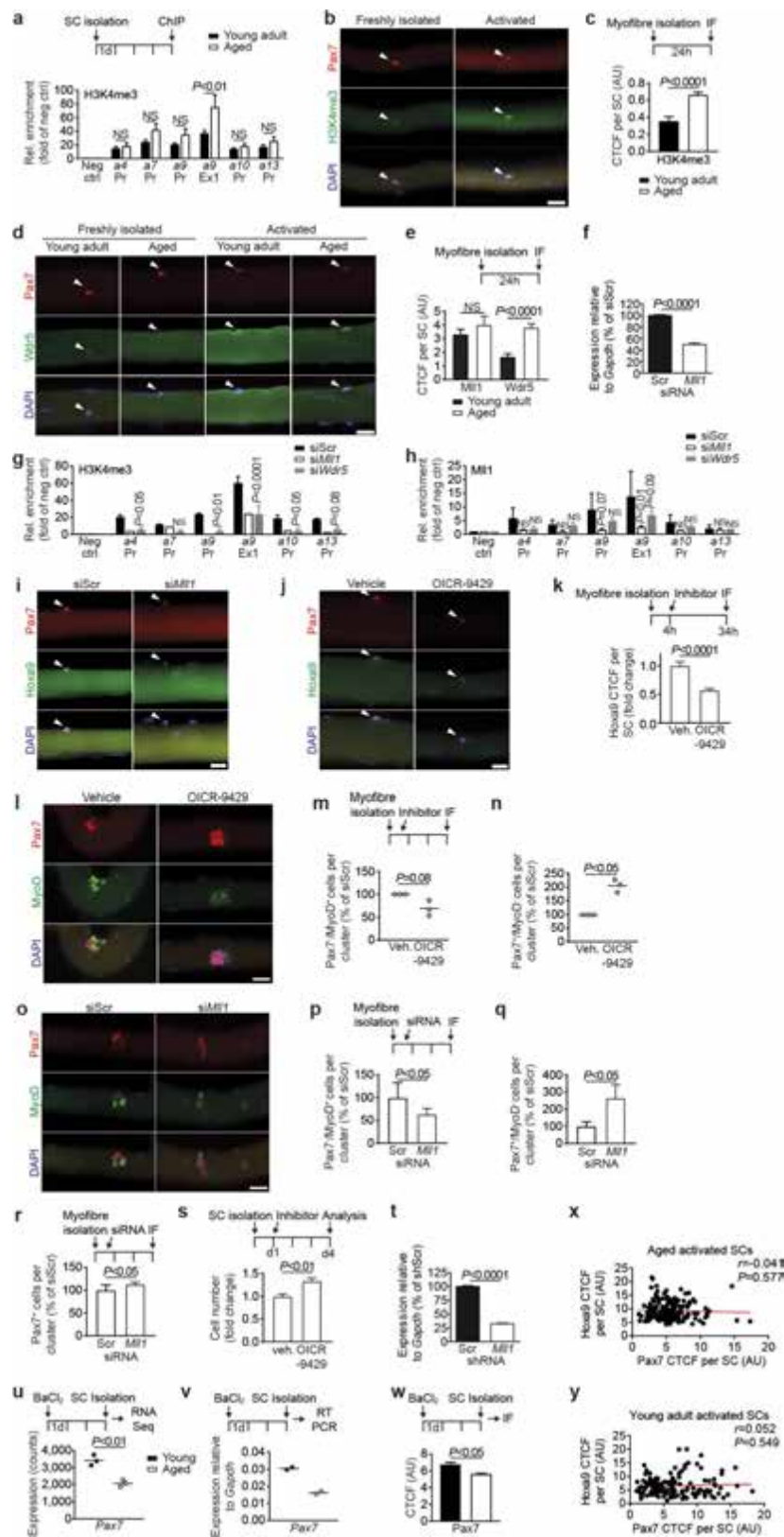


Extended Data Figure 5 | Inhibition of *Hoxa9* improves muscle regeneration in aged mice. **a**, Quantification of Pax7⁺ cells per area in uninjured tibialis anterior muscles from young adult and aged *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice. **b**, Representative immunofluorescence staining for Pax7 and laminin on tibialis anterior muscles from aged *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice that were collected 7 days after cardiotoxin (CTX) injury. **c**, qPCR analysis of *Hoxa9* expression in SCs isolated from tibialis anterior muscles injected with a self-delivering *Hoxa9* or scrambled siRNA and collected 5 days after muscle injury. **d**, Representative immunofluorescence staining for Pax7 and laminin of injured tibialis anterior muscles from young adult and aged mice that were injected with a self-delivery siRNA and collected 7 days after muscle injury. Nuclei were counterstained with DAPI (blue). Arrowheads denote Pax7⁺ cells. **e**, Quantification of Pax7⁺ cells from **d** per area. **f**, Frequency distribution minimal Feret's diameter of muscle fibres from **d**. **g**, Exemplary immunofluorescence staining for Pax7 and Ki67 on tibialis anterior muscles from aged *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice collected 7 days after muscle injury. Nuclei were counterstained with DAPI (blue). **h**, Quantification of proliferating SCs (Ki67⁺/Pax7⁺) as depicted in **g**. Scale bars, 50 μ m. *P* values were calculated by two-sided Student's *t*-test (**c**, **h**) or two-way ANOVA (**a**, **e**, **f**). *n* = 3 mice in **a**; *n* = 3 mice in **c**; *n* = 3 mice in **e**, **f**; *n* = 4 mice in **h**.



Extended Data Figure 6 | Inhibition of *Hoxa9* improves regenerative capacity of aged SCs. **a**, Quantification of donor-derived (eGFP⁺) myofibres from transplantation of SCs from young adult *Hoxa9*^{+/+} and *Hoxa9*^{-/-} mice. **b**, qPCR analysis of *Hoxa9* expression in SCs transduced with scrambled control or *Hoxa9* shRNA encoding lentivirus. **c–g**, Transplantation of eGFP-labelled SCs from young adult and aged mice that were targeted with shRNAs against *Hoxa9* or a scrambled control. **c**, Representative immunofluorescence staining for Pax7 and eGFP of transplanted muscle sections. Nuclei were counterstained with DAPI (blue). Arrowheads denote Pax7⁺/eGFP⁺ cells, asterisks label Pax7⁺/eGFP⁻ cells. **d**, Quantification of donor-derived (eGFP⁺) Pax7⁺ cells in **c**. **e**, Representative immunofluorescence staining for eGFP and laminin of transplanted muscle sections, nuclei were counterstained with DAPI (blue). **f**, **g**, Quantification of donor-derived (eGFP⁺) myofibres in **e** for two different *Hoxa9* shRNAs in two independent experiments.

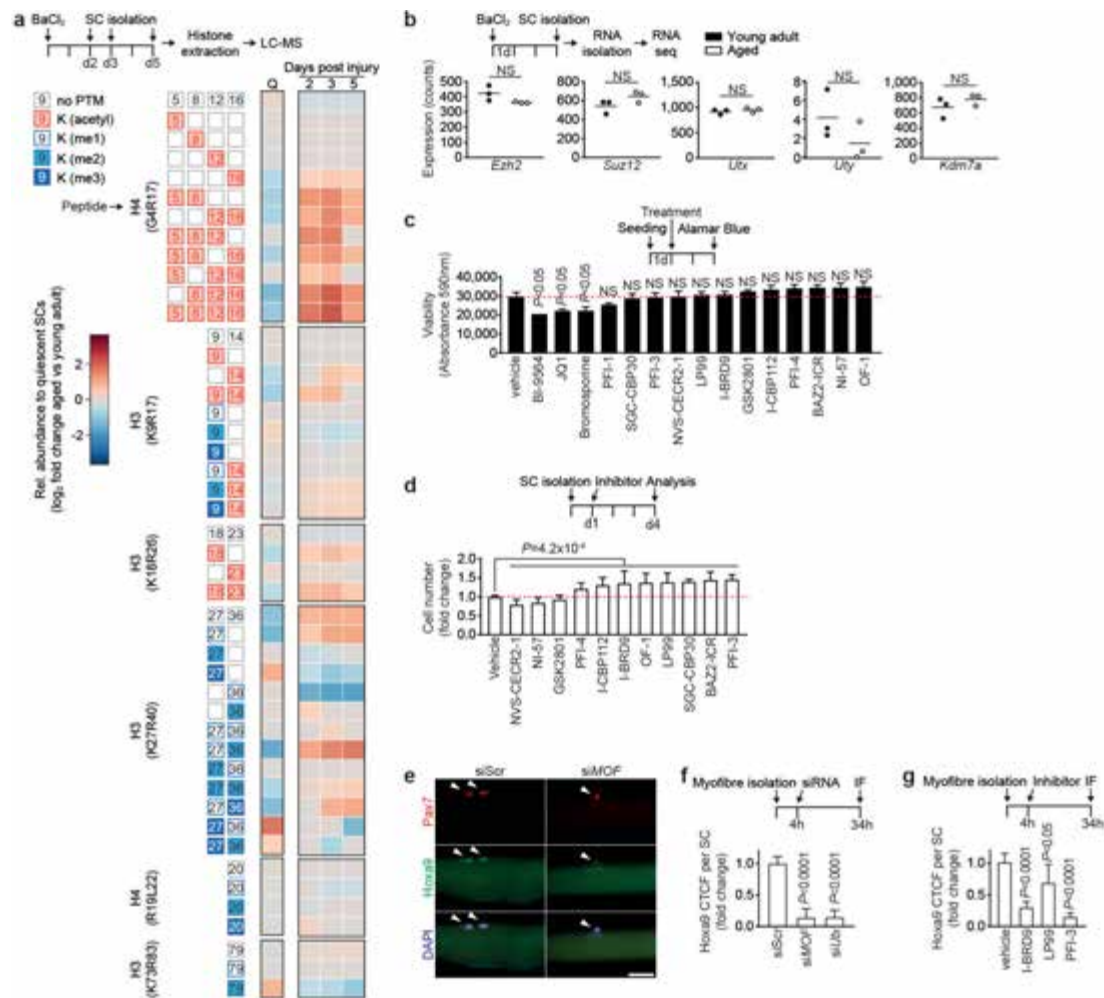
h, Exemplary immunofluorescence staining for eGFP and laminin in tibialis anterior muscles engrafted with untransduced aged SCs. Nuclei were counterstained with DAPI (blue). **i**, Flow cytometric analysis of transduction efficiency of donor SCs used for transplantation in primary recipients analysed in Fig. 2f. **j**, Representative flow cytometry plots for re-isolation of transplanted aged SCs that were untransduced as control or transduced with scrambled control or *Hoxa9* shRNA encoding lentivirus as quantified in Fig. 2f. **k**, Representative immunofluorescence staining for eGFP and laminin in engrafted tibialis anterior muscles from secondary recipients quantified in Fig. 2g. Nuclei were counterstained with DAPI (blue). Scale bars, 20 μ m (**c**), 50 μ m (**h**) and 100 μ m (**e**, **k**). *P* values were calculated by two-sided Student's *t*-test (**a**, **b**) or two-way ANOVA (**d**, **f**, **g**). *n* = 4 recipient mice in **a**; *n* = 3 mice in **b**; *n* = 6 recipient mice (young donors), *n* = 4 recipient mice (aged donors) in **d**, **f**; *n* = 5 recipient mice in **g**.



Extended Data Figure 7 | See next page for caption.

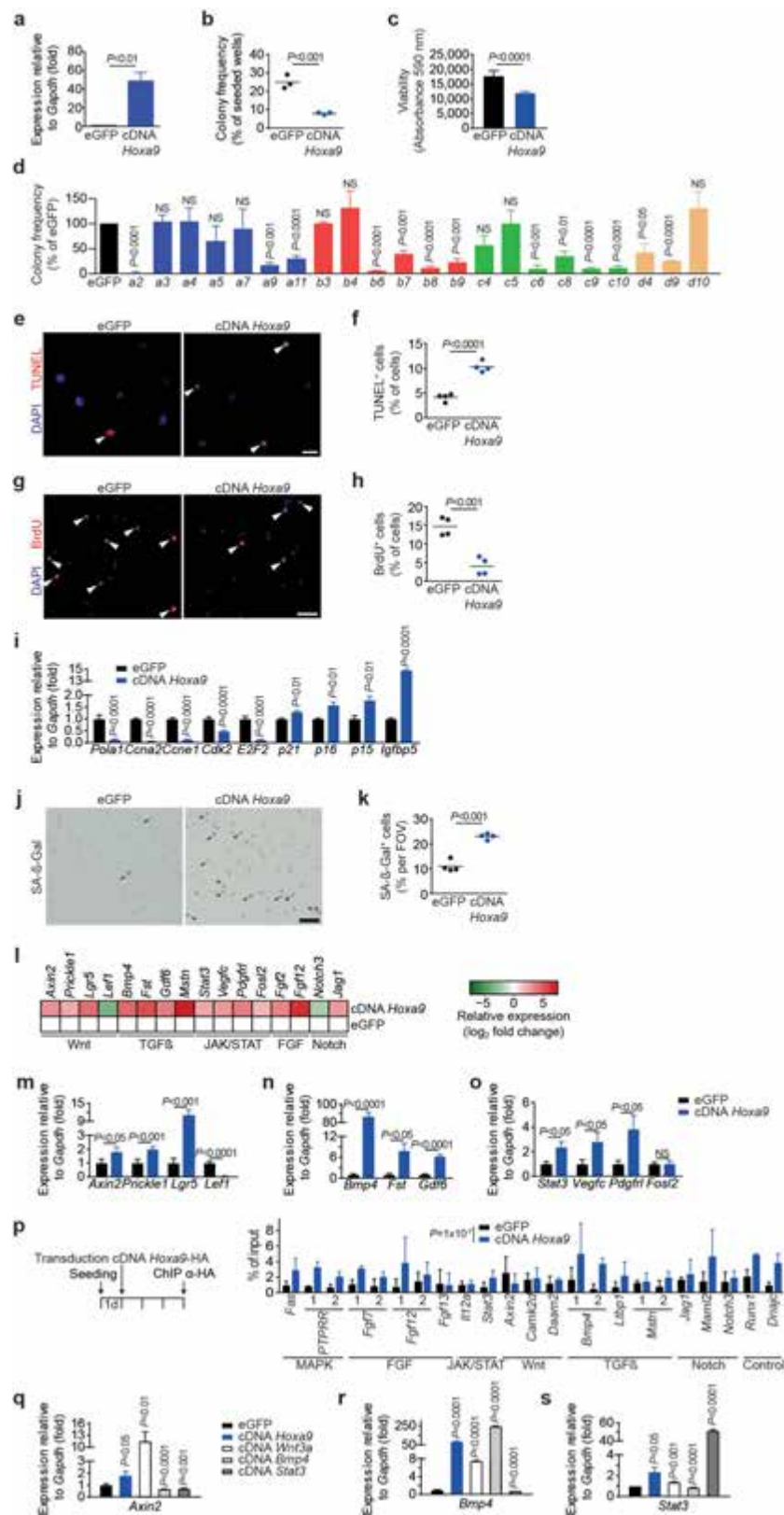
Extended Data Figure 7 | Inhibition of *Mll1* rescues H3K4me3 induction, *Hoxa9* overexpression, and functional impairment of activated SCs from aged mice. **a**, ChIP for H3K4me3 at promoters or exons of indicated Hox genes in activated SCs (4 day culture) from young adult and aged mice. **b**, Representative immunofluorescence staining for Pax7 and H3K4me3 on myofibre-associated SCs from aged mice that were freshly isolated or activated by 24-h culture of myofibres. **c**, Corrected total cell fluorescence (CTCF) for H3K4me3 on activated SCs shown in **b**. **d**, Representative immunofluorescence staining for Pax7 and Wdr5 on myofibre-associated SCs from young adult and aged mice that were freshly isolated or activated by 24-h culture of myofibres. **e**, CTCF for *Mll1* and Wdr5 per activated SC as shown in **d**. **f**, qPCR analysis of *Mll1* in SCs transfected with *Mll1* siRNA or scrambled control. **g**, **h**, ChIPs for H3K4me3 (**g**) and *Mll1* (**h**) in primary myoblasts 3 days after transfection with the indicated siRNAs. **i**, **j**, Immunofluorescence staining for Pax7 and *Hoxa9* in myofibres from aged mice after transfection with *Mll1* siRNA or scrambled control (**i**, quantification in Fig. 3d) or after treatment with OICR-9429 or vehicle (**j**). **k**, CTCF for *Hoxa9* per SC as shown in **j**. **l**, Representative immunofluorescence staining for Pax7 and MyoD on OICR-9429 treated myofibre-associated SCs from aged mice after 72 h culture. Nuclei were counterstained with DAPI (blue). **m**, **n**, Average number of Pax7⁻/MyoD⁺ cells (**m**) or Pax7⁺/MyoD⁻ cells (**n**) within clusters as shown in **l**. **o**, Representative immunofluorescence staining for Pax7 and MyoD on siRNA-treated myofibre-associated SCs from aged

mice after 72-h culture. Nuclei were counterstained with DAPI (blue). **p–r**, Average number of Pax7⁻/MyoD⁺ cells (**p**), Pax7⁺/MyoD⁻ cells (**q**) or Pax7⁺ cells (**r**) within clusters in **o**. **s**, Relative changes in cell number of aged SCs after treatment with OICR-9429 and 4 days of culture, compared to vehicle control. **t**, qPCR analysis of *Mll1* in SCs transduced with *Mll1* shRNA or scrambled control. **u–w**, Analysis of Pax7 expression in *in vivo* activated SCs from young adult and aged mice by RNA-sequencing (**u**), qPCR (**v**), or immunofluorescence as depicted in Fig. 1b (**w**). **x**, **y**, Pearson correlation comparing the *Hoxa9* immunofluorescence signal (quantification in Fig. 1c) and the Pax7 immunofluorescence signal (quantification in **w**) of activated SCs from aged (**x**) and young adult (**y**) mice. Note, there is no correlation between *Hoxa9* expression level and Pax7 expression level in activated SCs from aged mice. Scale bars, 20 μ m (**b**, **d**, **i**, **j**, **l**, **o**). *P* values were calculated by two-way ANOVA (**a**, **g**, **h**), two-sided Student's *t*-test (**f**, **m**, **n**, **p–v**), two-sided Mann–Whitney *U*-test (**c**, **e**, **k**, **w**) or Pearson correlation (**x**, **y**). *n* = 4 mice (young), *n* = 7 mice (aged) in **a**; *n* = 27 nuclei from 2 mice (young), *n* = 27 nuclei from 4 mice (aged) in **c**; *n* = 40/52 nuclei (*Mll1*), *n* = 44/99 nuclei (*Wdr5*) from 3 young/aged mice in **e**; *n* = 3 mice in **f**; *n* = 3 biological replicates (*Wdr5* siRNA), *n* = 2 biological replicates (*Mll1* siRNA) in **g**; *n* = 3 biological replicates in **h**; *n* = 173 nuclei (DMSO), *n* = 324 nuclei (OICR-9429) from 4 mice in **k**; *n* = 3 mice in **m**, **n**; *n* = 7 mice in **p–r**; *n* = 6 mice in **s**; *n* = 3 mice in **t**; *n* = 3 mice in **u**; *n* = 2 mice in **v**; *n* = 134 nuclei (young), *n* = 181 nuclei (aged) from 3 mice in **w–y**.



Extended Data Figure 8 | Alterations in the epigenetic stress response of activated SCs from aged mice. **a**, Heatmap displaying relative changes in abundance of different histone modifications (measured at the indicated peptides) in freshly isolated SCs from aged compared to young adult mice. SCs were analysed in quiescence (Q, derived from uninjured muscle) or at the indicated time points after activation mediated by muscle injury. Relative abundances at indicated days after injury are first normalized to quiescent SCs, and then compared between SCs isolated from aged and young adult mice and log₂ scaled. Only significant changes are shown ($P < 0.05$). **b**, Expression analysis of the indicated genes in freshly isolated *in vivo* activated SCs from young adult and aged mice based on RNA-sequencing. **c**, Viability of primary myoblasts after 48-h treatment with bromodomain inhibitors (1 μ M) from the Structural Genomics Consortium probe set, measured by Alamar Blue assay. **d**, Relative changes in cell number of aged SCs after treatment with non-toxic bromodomain

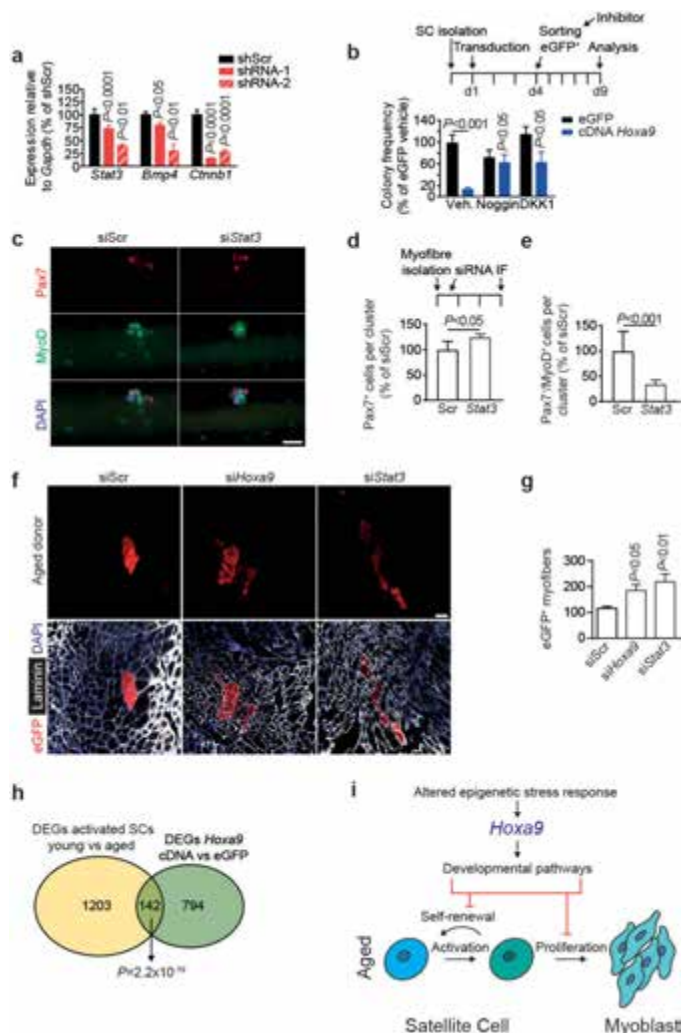
inhibitors (1 μ M) from **c** and 4 days of culture, compared to vehicle control. A Wilcoxon rank-sum test on the ratio of all cell counts being equal to 1 was performed to test the hypothesis of a general effect of the inhibitors on cell number. **e**, Representative immunofluorescence staining for Pax7 and Hoxa9 in siRNA-treated myofibre-associated SCs from aged mice. Scale bar, 20 μ m. **f**, CTCTF for Hoxa9 per SC as shown in **e**. **g**, Quantification of immunofluorescence staining for Hoxa9 in Pax7⁺ cells on myofibre-associated SCs from aged mice treated with bromodomain inhibitors. P values were calculated by two-sided Student's t -test (**a–c**), Wilcoxon rank-sum test (**d**) or two-sided Mann–Whitney U -test (**f, g**). $n = 4$ mice in **a**; $n = 3$ mice in **b**; $n = 4$ biological replicates in **c**; $n = 6$ mice in **d**; $n = 71$ nuclei (scrambled siRNA), $n = 48$ nuclei (*MOF* siRNA), $n = 98$ nuclei (*Utx* siRNA) from 3 mice in **f**; $n = 60$ nuclei (vehicle), $n = 59$ nuclei (*I-BRD9*), $n = 38$ nuclei (*LP99*), $n = 62$ nuclei (*PFI-3*) from 3 mice in **g**.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Overexpression of *Hox* genes inhibits SC function. **a**, Expression of *Hoxa9* in SCs transduced with *Hoxa9* cDNA or eGFP as control. **b**, **c**, FACS-isolated SCs from young adult mice were transduced with a lentivirus either containing both eGFP and *Hoxa9* cDNA or only eGFP. Infected (eGFP⁺) cells were isolated after 3 days. **b**, Frequency of myogenic colonies from single-cell-sorted SCs. **c**, Quantification of cell number based on Alamar Blue assay of bulk cultures. **d**, Frequency of myogenic colonies of SCs overexpressing the indicated *Hox* genes. **e**, **g**, TUNEL (**e**) or BrdU (**g**) staining of SCs overexpressing *Hoxa9* or eGFP. Infected (eGFP⁺) cells were isolated 3 days after transduction and analysed 3 days later. Nuclei were counterstained with DAPI (blue). Arrowheads mark TUNEL- or BrdU-positive cells. **f**, **h**, Quantification of apoptosis (**f**) or proliferation (**h**) based on TUNEL or BrdU staining as in **e** or **g**. **i**, qPCR-based expression analysis of various cell-cycle and senescence markers in SCs overexpressing *Hoxa9* compared to eGFP-infected controls, 5 days after infection. **j**, Senescence-associated- β -galactosidase (SA- β -Gal) staining of SCs overexpressing *Hoxa9* or eGFP at day 5 after infection. Arrowheads mark SA- β -Gal-positive cells. **k**, Quantification of senescence per field of view (FOV) based on SA- β -Gal staining in **j**. **l**, Heatmap displaying log₂ fold changes

of expression of selected genes from microarray analysis in Fig. 5a. **m–o**, qPCR validation of differentially expressed genes annotated to Wnt (**m**), TGF β (**n**) and JAK/STAT pathways (**o**) as in **l**. **p**, Identification of *Hoxa9*-binding sites by anti-HA ChIP of primary myoblasts overexpressing HA-tagged *Hoxa9* cDNA or eGFP as control. Shown is the qPCR for 1 or 2 putative *Hoxa9*-binding sites at the indicated loci. *Hoxa9*-binding sites at target genes were identified as described in the Methods and are listed in Supplementary Table 1. A two-sided block bootstrap test on the difference of the percentage of bound DNA for all binding sites being equal to 0 was performed to test the hypothesis of a generally increased binding of *Hoxa9*. **q–s**, SCs were infected with lentiviruses expressing *Hoxa9*, *Wnt3a*, *Bmp4* or *Stat3* cDNAs or eGFP. qPCR analysis of expression of the indicated target genes at 5 days after infection: *Axin2* (**q**), *Bmp4* (**r**) and *Stat3* (**s**). Scale bars, 20 μ m (**e**, **g**) and 50 μ m (**j**). *P* values were calculated by two-sided Student's *t*-test (**a–d**, **f**, **h**, **k**, **q–s**) or two-way ANOVA (**i**, **m–o**). *n* = 4 mice in **a**; *n* = 3 mice in **b**; *n* = 7 mice in **c**; *n* = 3 mice in **d**; *n* = 4 mice in **f**, **h**, **k**; *n* = 3 mice (p15, p21), *n* = 6 mice (p16), *n* = 4 mice (all others) in **i**; *n* = 4 pools of 3 mice in **l**; *n* = 4 mice in **m–o**; *n* = 3 biological replicates for **p**; *n* = 3 mice (*Wnt3a*, *Bmp4*, *Stat3*), *n* = 4 mice (eGFP, *Hoxa9*) in **q–s**.



Extended Data Figure 10 | Validation of *Hoxa9* downstream targets.

a, Knockdown efficiency of two shRNAs (red bars) for *Stat3*, *Bmp4* and *Ctnnb1*. **b**, SCs from young adult mice were transduced with an *Hoxa9* and *eGFP*-encoding lentivirus. *eGFP*⁺ cells were sorted as single cells and cultured in the presence of noggin, DKK1 or 0.1% BSA in PBS as vehicle. Colony frequency was assessed after 5 days and is compared to *Hoxa9* cDNA expressing cells treated with vehicle control. **c**, Representative immunofluorescence staining for Pax7 and MyoD on siRNA-transfected myofibers from aged mice after 72 h of culture. Nuclei were counterstained with DAPI (blue). **d**, **e**, Average number of Pax7⁺ cells (**d**) or Pax7⁺/MyoD⁺ cells (**e**) within clusters in **c**. **f**, Representative immunofluorescence staining for eGFP and laminin in tibialis anterior muscles engrafted with siRNA-transfected SCs isolated from eGFP transgenic aged mice. Nuclei were counterstained with DAPI (blue). **g**, Quantification of donor-derived (*eGFP*⁺) myofibers in **f**. **h**, Area-proportional Venn diagram of differentially expressed genes from indicated transcriptomes. **i**, Model for the *Hoxa9*-mediated impairment of SC function during ageing: quiescent SCs become activated upon muscle injury and proliferate as myoblasts to repair damaged muscle tissue. After activation, aged SCs display global and locus-specific alterations in the epigenetic stress response resulting in overexpression of *Hoxa9*, which in turn induces developmental pathways inhibiting SC function and muscle regeneration in aged mice. Scale bars, 20 μ m (**c**), and 100 μ m (**f**). *P* values were calculated by two-way ANOVA (**a**, **b**) or two-sided Student's *t*-test (**d**, **e**, **g**). *n* = 3 mice in **a**; *n* = 4 mice in **b**; *n* = 5 mice in **d**, **e**; *n* = 5 recipient mice in **g**; *n* = 3 mice per group (activated SCs), *n* = 4 pools of 3 mice (*Hoxa9* overexpression) in **h**.

A 17-gene stemness score for rapid determination of risk in acute leukaemia

Stanley W. K. Ng^{1*}, Amanda Mitchell^{2*}, James A. Kennedy^{2,3,4*}, Weihsu C. Chen², Jessica McLeod², Narmin Ibrahimova², Andrea Arruda², Andreea Popescu², Vikas Gupta^{2,3,4}, Aaron D. Schimmer^{2,3,4,5}, Andre C. Schuh^{2,3,4}, Karen W. Yee^{2,3,4}, Lars Bullinger⁶, Tobias Herold^{7,8}, Dennis Görlich⁹, Thomas Büchner^{10,†}, Wolfgang Hiddemann^{7,8}, Wolfgang E. Berdel¹⁰, Bernhard Wörmann¹¹, Meyling Cheok¹², Claude Preudhomme¹³, Hervé Dombret¹⁴, Klaus Metzeler^{7,8}, Christian Buske¹⁵, Bob Löwenberg¹⁶, Peter J. M. Valk¹⁶, Peter W. Zandstra¹, Mark D. Minden^{2,3,4,5,§}, John E. Dick^{2,17,§} & Jean C. Y. Wang^{2,3,4,§}

Refractoriness to induction chemotherapy and relapse after achievement of remission are the main obstacles to cure in acute myeloid leukaemia (AML)¹. After standard induction chemotherapy, patients are assigned to different post-remission strategies on the basis of cytogenetic and molecular abnormalities that broadly define adverse, intermediate and favourable risk categories^{2,3}. However, some patients do not respond to induction therapy and another subset will eventually relapse despite the lack of adverse risk factors⁴. There is an urgent need for better biomarkers to identify these high-risk patients before starting induction chemotherapy, to enable testing of alternative induction strategies in clinical trials⁵. The high rate of relapse in AML has been attributed to the persistence of leukaemia stem cells (LSCs), which possess a number of stem cell properties, including quiescence, that are linked to therapy resistance^{6–10}. Here, to develop predictive and/or prognostic biomarkers related to stemness, we generated a list of genes that are differentially expressed between 138 LSC⁺ and 89 LSC[−] cell fractions from 78 AML patients validated by xenotransplantation. To extract the core transcriptional components of stemness relevant to clinical outcomes, we performed sparse regression analysis of LSC gene expression against survival in a large training cohort, generating a 17-gene LSC score (LSC17). The LSC17 score was highly prognostic in five independent cohorts comprising patients of diverse AML subtypes ($n = 908$) and contributed greatly to accurate prediction of initial therapy resistance. Patients with high LSC17 scores had poor outcomes with current treatments including allogeneic stem cell transplantation. The LSC17 score provides clinicians with a rapid and powerful tool to identify AML patients who do not benefit from standard therapy and who should be enrolled in trials evaluating novel upfront or post-remission strategies.

To derive an LSC-based biomarker, 83 cell samples obtained from 78 AML patients (Extended Data Fig. 1a) were sorted into fractions based on expression of CD34 and CD38, and LSC activity in each fraction was assessed by xenotransplantation into NOD.*Prkdc^{scid}.Il2r^{gnull}* (NSG) mice (Extended Data Fig. 1b). Consistent with previous reports, the majority of CD34⁺ and a minority of CD34[−] fractions contained

LSCs^{11,12}. However, LSCs were detected in fractions of all CD34/CD38 phenotypes (Extended Data Fig. 1c, d), underscoring the importance of performing functional assays to define LSC activity.

Each of the functionally defined 138 LSC⁺ and 89 LSC[−] fractions was subjected to gene expression (GE) analysis. By comparing GE profiles of LSC⁺ and LSC[−] fractions, a list of differentially expressed (DE) genes was obtained; 104 genes exhibited ≥ 2 -fold expression level differences ($P < 0.01$; Extended Data Fig. 1e and Extended Data Table 1). We defined an LSC⁺ reference profile as the average expression levels of these 104 genes in the LSC⁺ fractions. There was a strong correlation between engraftment ability of individual cell fractions and their GE similarity to the LSC⁺ reference profile, as well as to the global GE profiles of normal haematopoietic stem cells (HSCs) and multipotent progenitors (MPPs) from human umbilical cord blood¹³ (Fig. 1a, b). Conversely, GE similarity to the LSC⁺ reference profile was anti-correlated with the global GE patterns of mature myeloid cell types, including granulocytes and monocytes¹⁴ (Fig. 1b). These findings suggest that the 104 genes most DE between LSC⁺ and LSC[−] cell populations are associated with stem cell transcriptional programs that are shared between LSC and normal HSCs/MPPs.

To extract the core transcriptional components of stemness that relate to clinical outcomes across a broad spectrum of AML patient subtypes, we interrogated a large data set of 495 patients (Gene Expression Omnibus (GEO) accession GSE6891 (ref. 15)), in which 89 of the 104 DE LSC genes were captured. Expression of the 89 genes in these unfractionated patient samples was variable and showed a similar pattern of correlation to the LSC⁺ reference profile, as did the sorted LSC⁺ and LSC[−] fractions (Fig. 1c), suggesting that LSC-associated GE programs are detectable at the bulk cell level. We applied a statistical regression algorithm based on the least absolute shrinkage and selection operator (LASSO)^{16,17} to relate GE to patient survival in this training cohort, using either the full list of 89 LSC genes or the subset of 43 genes more highly expressed in LSC⁺ fractions. Analysis of the latter subset yielded an optimal 17-gene signature (LSC17 score), which could be calculated for each patient as the weighted sum of expression of the 17 genes (Fig. 1c). High LSC17 scores were strongly associated with poor overall survival (OS) and event-free survival (EFS) (Extended

¹Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario M5G 1A1, Canada. ²Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 2M9, Canada. ³Division of Medical Oncology and Hematology, Department of Medicine, University Health Network, Toronto, Ontario M5G 2M9, Canada. ⁴Department of Medicine, University of Toronto, Toronto, Ontario M5G 1A1, Canada. ⁵Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1A1, Canada. ⁶Department of Internal Medicine III, University Hospital of Ulm, 89081 Ulm, Germany. ⁷Department of Internal Medicine III, University of Munich, 81377 Munich, Germany. ⁸German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. ⁹Institute of Biostatistics and Clinical Research, University of Münster, 48149 Münster, Germany. ¹⁰Department of Medicine, Hematology and Oncology, University of Münster, 48149 Münster, Germany. ¹¹Department of Hematology, Oncology and Tumor Immunology, Charité University Medicine, Campus Virchow, 10117 Berlin, Germany. ¹²Jean-Pierre AUBERT Research Center UMR-S 1172, Institute for Cancer Research Lille, 59045 Lille, France. ¹³University Hospital of Lille, Center of Pathology, Laboratory of Hematology, 59037 Lille, France. ¹⁴Saint-Louis Hospital, Department of Hematology, University of Paris Diderot, 75010 Paris, France. ¹⁵Comprehensive Cancer Center Ulm, Institute of Experimental Cancer Research, University Hospital of Ulm, 89081 Ulm, Germany. ¹⁶Department of Hematology, Erasmus University Medical Centre, 3015 CE Rotterdam, the Netherlands. ¹⁷Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5G 1A1, Canada.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

†Deceased.

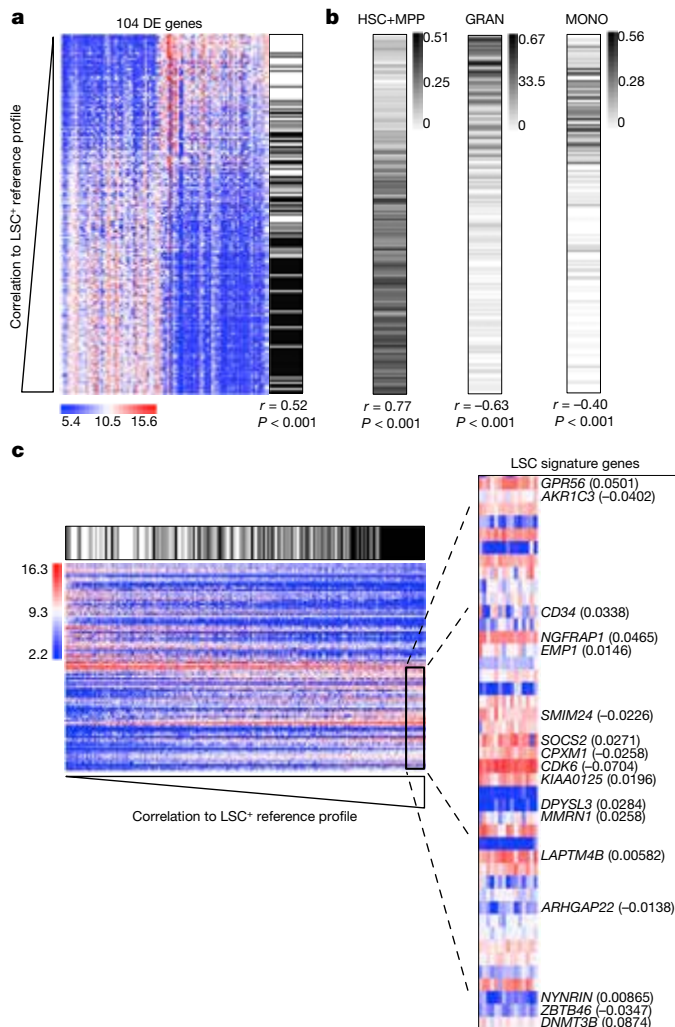


Figure 1 | Analysis of LSC-specific GE identifies an optimal 17-gene prognostic signature. **a**, GE patterns of the top 104 genes (columns) DE between 138 LSC⁺ and 89 LSC⁻ cell fractions (rows). Horizontal black and white bars denote LSC⁺ and LSC⁻ fractions, respectively. r = correlation coefficient between engraftment status and similarity to the LSC⁺ reference profile. **b**, Similarity of global GE of each cell fraction to that of stem-cell-enriched cell populations (HSCs plus MPPs) and mature myeloid cell populations (granulocytes, GRAN; monocytes, MONO) from human umbilical cord blood. r = correlation coefficient between similarity to the LSC⁺ reference profile and similarity to the cell types indicated. **c**, GE patterns of 89/104 DE genes captured in the GSE6891 data set. The relative ordering of the 89 genes is the same as in **b** (rotated counterclockwise 90°). Vertical black and white bars denote samples with LSC17 scores above and below the median, respectively. The 17 signature genes are depicted in the magnified view on the right (regression coefficients in parentheses).

Data Fig. 1f). In addition, patients with high LSC17 scores had significantly higher percentages of bone marrow blasts at diagnosis, a higher incidence of the *FLT3* internal tandem duplication mutation (*FLT3*-ITD) and adverse cytogenetics, higher rates of relapse, and lower response rates to standard induction chemotherapy, reflecting a link between LSC-associated GE programs and clinical outcomes.

We evaluated the association of the LSC17 score with survival in three independent AML cohorts (one from The Cancer Genome Atlas (TCGA)¹⁸, two from GEO accession GSE12417 (ref. 19)). In the TCGA AML cohort ($n = 183$), patients with a high LSC17 score had significantly shorter OS than patients with a low score (Fig. 2a and Extended Data Table 2; hazard ratio (HR) = 2.62; $P < 0.001$). This survival difference was also found for the subset of cytogenetically normal (CN)-AML patients ($n = 83$) (Extended Data Fig. 2a; median

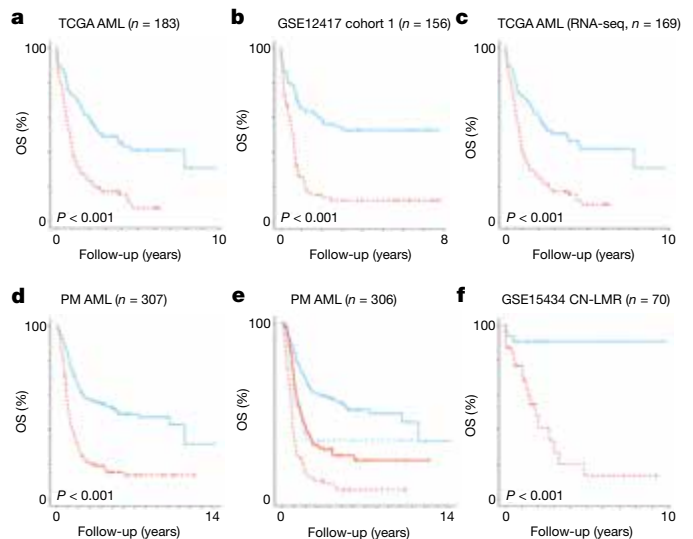


Figure 2 | LSC signature scores are associated with OS in multiple independent AML cohorts across different GE measurement platforms. **a–d**, Kaplan–Meier estimates of OS, according to LSC17 scores calculated using microarray (**a**, **b**), RNA-seq (**c**), or NanoString (**d**) GE data sets. **e**, Kaplan–Meier estimates of OS according to LSC17 score and whether or not CR was achieved after initial therapy (no CR, dotted lines; CR, solid lines). **f**, Kaplan–Meier estimates of OS, according to LSC3 scores calculated using microarray GE data. For all panels, OS of patients with scores above and below the median in each cohort are shown by red and blue lines, respectively.

OS 10.4 versus 24.1 months; HR = 2.06; $P = 0.006$). Similar results were observed in the two CN-AML cohorts from GSE12417 (cohort 1: Fig. 2b, HR = 3.16, $P < 0.001$; cohort 2: Extended Data Fig. 2b, HR = 2.66, $P = 0.002$; Extended Data Table 3). In GSE12417 cohort 1, a high LSC17 score was associated with shorter OS regardless of whether or not complete remission (CR) was achieved (Extended Data Fig. 2c; CR: median OS 9.8 months versus not reached; HR = 3.20; $P < 0.001$; no CR: median OS 2.0 versus 2.5 months; HR = 1.58; $P = 0.14$). As in the training cohort, high LSC17 scores were significantly associated with adverse cytogenetic and molecular features, failure to achieve CR, and shorter EFS and relapse-free survival (RFS) (Extended Data Fig. 2d–g and Extended Data Tables 2, 3). When applied to RNA-sequencing (RNA-seq) data for the TCGA cohort, the LSC17 score remained highly associated with outcome (full cohort: Fig. 2c, HR = 2.48, $P < 0.001$; CN-AML subset: Extended Data Fig. 2h, HR = 2.38, $P = 0.001$), demonstrating robustness across technology platforms. In multivariate survival analysis using Cox proportional hazards (CPH) models, the LSC17 score retained significant prognostic value in all tested cohorts independent of known predictors of outcome including patient age, presenting white blood cell (WBC) count, cytogenetic risk group, type of AML (*de novo* versus secondary), and the presence of *FLT3*-ITD and *NPM1* mutations (Extended Data Table 4a).

Recent studies mapping the mutational landscape of AML have identified additional recurrent mutations that carry independent prognostic information²⁰. Of our validation cohorts, extensive mutational profiling data was available only for the TCGA AML cohort. In this data set, six mutations frequently found in AML occurred in at least three patients and were also significantly associated with OS as single factors in univariate survival analysis. However, in a multivariate CPH model that included all six of these mutations as well as common clinical parameters (age, WBC count, cytogenetic risk group), only *DNMT3A* retained prognostic significance when the LSC17 score was included in the model, whereas the LSC17 score remained a strong and significant independent prognostic factor (Extended Data Table 5a). Recently, a comprehensive genomic classification scheme was reported and was shown to be more accurate for patient risk stratification than

the European LeukaemiaNet risk group definitions^{3,21}. When this new scheme was applied to the TCGA AML cohort, inclusion of the LSC17 score in multivariate CPH models significantly improved the overall strength of association of the model with patient OS (Extended Data Table 5b, $P < 0.001$, likelihood ratio test), and the LSC17 score itself remained statistically significant. Three of the fourteen subgroups in the new genomic classification scheme are less well characterized ('driver mutations but not class-defining', 'no detected driver mutations', and 'meeting criteria for 2 or more subgroups'); patients in these groups had similar survival. The LSC17 score was able to discriminate between shorter and longer OS in the combined subset of patients falling into these three subgroups, and thus refines this state-of-the-art genomics classification scheme (Extended Data Fig. 2i and Extended Data Table 5c).

The LSC17 score displayed superior prognostic accuracy when tested against other published LSC signatures derived from GE analysis of cell populations defined phenotypically or by multidimensional mass cytometry^{22,23}, or generated based on epigenetic differences between a small number of functionally tested cell populations²⁴. These other signatures, like our previously reported 42-gene LSC signature¹¹, are lists of DE genes generated by comparing cellular phenotypes or functional states without further statistical analysis of the contribution of each gene to explaining patient outcome. When tested in three independent cohorts (GSE12417 CN-AML cohorts 1 and 2, and TCGA AML), these other signatures were prognostic in some cases as single factors or when controlling for common clinical covariates. However, when the LSC17 score was incorporated in multivariate analysis, they were no longer significantly associated with survival, whereas the LSC17 score remained highly prognostic (Extended Data Table 6).

To develop a clinically applicable GE-based diagnostic test, we turned to the NanoString platform, which is reproducible, cost effective, and has a rapid turnaround time of 24–48 h (ref. 25). We designed a custom NanoString assay and generated GE data for 307 AML patients treated at the Princess Margaret (PM) Cancer Centre. A high LSC17 score was associated with known adverse prognostic features including older age, high initial WBC count, and unfavourable cytogenetics (Extended Data Table 7a). Consistent with our findings using microarray and RNA-seq GE data, patients with high LSC17 scores had significantly shorter OS than patients with low scores (Fig. 2d and Extended Data Table 7a; HR = 2.73; $P < 0.001$); this was true regardless of whether or not remission was achieved after primary induction therapy (Fig. 2e; CR: median OS 18.9 versus 90.3 months; HR = 2.18; $P < 0.001$; no CR: median OS 10.5 versus 20.7 months; HR = 2.16; $P = 0.02$). Similarly, a high LSC17 score was associated with shorter EFS and RFS (Extended Data Fig. 2j–m). The association between a high LSC17 score and shorter OS was also observed in the subset of patients with CN-AML (Extended Data Fig. 2n; median OS 13.7 versus 65.7 months; HR = 2.64; $P < 0.001$). Importantly, in multivariate survival analysis including established risk factors, the LSC17 score retained independent prognostic value in both the full cohort as well as in the CN-AML subset (Extended Data Table 4a). Together, these results demonstrate the broad applicability and strong prognostic value of the LSC17 score on the clinically serviceable NanoString platform.

Allogeneic stem cell transplantation (aSCT) has strong anti-leukaemic effects; however, potential benefits can be offset by considerable transplant-related mortality and thus the procedure is generally reserved for patients with a higher risk of relapse on the basis of available risk features (for example, cytogenetics, assessment of minimal residual disease)²⁶. Inclusion of aSCT as a time-dependent covariate in the PM AML cohort (univariate Mantel–Byar analysis; Fig. 3a) did not demonstrate a significant impact of aSCT on OS for either high- or low-score patients (high LSC17 score, $P = 0.20$; low LSC17 score, $P = 0.06$). Furthermore, a high LSC17 score was associated with shorter OS, irrespective of whether or not patients underwent aSCT (aSCT: median OS 11.7 versus 28.4 months for high versus low LSC17 score, respectively; HR = 2.14; $P = 0.005$; no aSCT: median OS

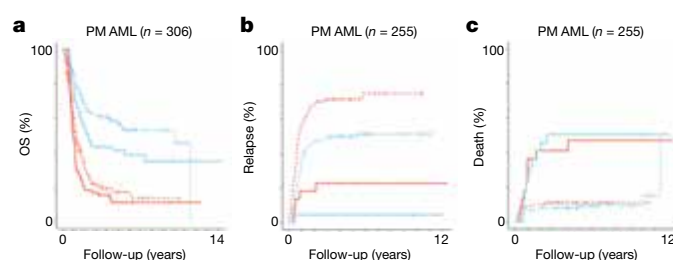


Figure 3 | Impact of aSCT on patient outcome. **a**, Simon and Makuch estimates of OS, according to LSC17 scores computed using NanoString GE data and whether or not patients received aSCT, **b**, **c**, Time from CR1 to first relapse (**b**) or death (**c**) as competing risks, as estimated by cumulative incidence analysis, according to LSC17 scores calculated using NanoString GE data. In all panels, red and blue lines show patients with scores above and below the median in each set, respectively, while solid and dotted lines denote patients who did and did not undergo aSCT, respectively.

14.7 versus 123.3 months; HR = 2.99; $P < 0.001$). The LSC17 score retained prognostic value when adjusted for common clinical factors in multivariate Andersen–Gill models (aSCT: HR = 2.00, $P = 0.04$; no aSCT: HR = 2.63, $P < 0.001$). Similar results were observed in the subset of CN-AML cases (Extended Data Fig. 2o), and in the analysis of EFS and RFS (data not shown).

We also examined the cumulative incidence of the competing risks of relapse and death from time of first CR (CR1) in patients who did or did not undergo aSCT in the PM AML cohort. A high LSC17 score was associated with earlier relapse in the subset of patients who did not undergo aSCT, in both univariate (Fig. 3b, sub-distribution HR (SHR) = 1.92; Gray's test $P < 0.001$; median time to relapse 9.31 versus 65.2 months) and Fine–Gray multivariate analysis (SHR = 1.85, $P = 0.003$). aSCT reduced the risk of relapse, although small patient numbers precluded seeing a statistically significant difference between high- and low-score patients (SHR = 5.26; $P = 0.09$). However, the reduction in relapse risk was offset by a significantly greater risk of death in both high- and low-score groups compared to patients who did not undergo aSCT (Fig. 3c; $P < 0.001$). Indeed, the risk of death after aSCT versus risk of relapse without aSCT was very similar for low-score patients. Thus, the LSC17 score will aid in defining which patients should undergo aSCT.

The LSC17 score was initially trained using clinical data from the GSE6891 data set, which included only a small group ($n = 44/495$, 9%) of CN-AML patients classified as low molecular risk (CN-LMR, defined as the presence of *NPM1* mutation and no *FLT3*-ITD); as such, survival differences within this small patient subset might not have been captured optimally by the statistical regression algorithm applied to the entire cohort. We therefore retrained the 17 LSC signature genes against OS for only the CN-LMR cases in GSE6891 and identified an optimized, reweighted sub-signature in which only 3 of the 17 genes contributed to the calculated score (LSC3). A high LSC3 score identified patients with poor outcome in an independent cohort of CN-LMR patients (GEO accession GSE15434 (ref. 27)) (Fig. 2f and Extended Data Table 7b; HR = 8.41; $P < 0.001$), and was strongly associated with shorter survival in the subset of 29 CN-LMR cases in the PM cohort analysed by NanoString methodology (Extended Data Fig. 2p; median OS 39.2 months versus not reached, HR = 3.65, $P = 0.05$), retaining independent prognostic value in multivariate analysis (Extended Data Table 4b). These findings demonstrate the feasibility of optimizing the LSC17 score for selected patient subsets.

We next tested the ability of the LSC17 score to predict therapy resistance (defined as failure to achieve CR after initial induction)⁵, as this remains one of the primary barriers to cure. In the PM AML cohort, the LSC17 score as a single continuous variable was more predictive of therapy resistance than cytogenetic risk (area under the receiver operating characteristic curve (AUROC) = 0.78 versus 0.70). In multivariate logistic regression models that also considered age, WBC

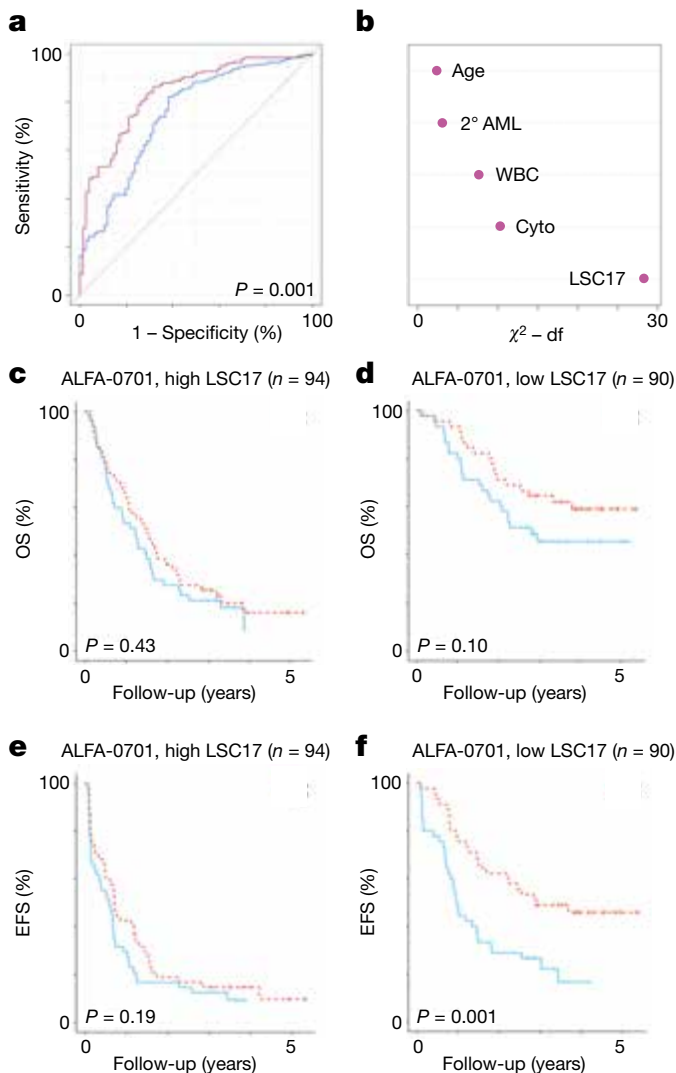


Figure 4 | LSC17 score predicts therapy response. **a**, Receiver operating characteristic (ROC) curves for prediction of initial therapy resistance, using logistic regression models that include age, WBC count, cytogenetic risk (Cyto), and *de novo* versus secondary AML (2° AML) as covariates, with (maroon line) or without (blue line) LSC17 score. **b**, Significance (chi-squared statistic) of each covariate for prediction of therapy resistance in the multivariate model that includes LSC17 scores. df, degrees of freedom. **c–f**, Kaplan–Meier estimates of OS (**c**, **d**) or EFS (**e**, **f**) for patients with high (**c**, **e**) or low (**d**, **f**) LSC17 scores treated with standard chemotherapy with (red lines) or without (blue lines) addition of GO.

count, cytogenetic risk and *de novo* versus secondary AML, inclusion of the LSC17 score markedly improved predictive ability (Fig. 4a; AUROC = 0.82 versus 0.73, increased sensitivity = 3.38%, increased specificity = 9.20%), and LSC17 score was the most significant covariate as measured by the Wald chi-squared statistic (Fig. 4b). Multivariate models that included either cytogenetic risk or continuous LSC17 score had comparable predictive value for therapy resistance (with LSC17: AUROC = 0.79 versus 0.73, increased sensitivity = 2.10%, increased specificity = 5.71%, $P = 0.12$). As the LSC17 score was trained to associate with OS, we tested whether reweighting the 17 genes to predict treatment response directly would result in even stronger predictive ability, using a random 50:50 split of the PM cohort for training and testing. Indeed, the retrained response score had better predictive value as a single factor than the unadjusted LSC17 score (AUROC = 0.81 versus 0.78). These results demonstrate that the LSC17 score improves the ability to predict therapy resistance in newly diagnosed AML patients.

We also used a data set from the ALFA-0701 trial^{28,29} (Extended Data Table 8a) to test the ability of the LSC17 score to predict response to gemtuzumab ozogamicin (GO), a drug–antibody conjugate shown to improve survival when added to standard induction chemotherapy. A higher LSC17 score was associated with shorter OS irrespective of treatment arm (Extended Data Fig. 2q; median OS 15.4 versus 46.2 months; HR = 2.45; $P < 0.001$). Notably, patients with low but not high LSC17 scores benefited from addition of GO to standard chemotherapy, with longer OS, EFS and RFS (OS: Fig. 4c, d, median not reached versus 34.3 months, HR = 0.60, $P = 0.11$; EFS: Fig. 4e, f, median 35.4 versus 11.7 months, HR = 0.42, $P = 0.001$; RFS: Extended Data Fig. 2r, s, median not reached versus 16.4 months, HR = 0.53, $P = 0.03$; Extended Data Table 8b). These data suggest that the LSC17 score could be used to facilitate more rational use of GO in patients most likely to benefit, while sparing high-score patients who do not derive benefit any potential toxicities.

Many cancer biomarkers rely on mutational profiling. However, the high degree of molecular complexity in AML presents a considerable challenge to clinical implementation of such approaches^{18,21,30}. The strong prognostic value of the LSC17 score across the spectrum of AML genotypes suggests that perturbations caused by driver mutations coalesce on alterations in stemness properties, and that the LSC17 score is able to distil these downstream consequences. A high LSC17 score probably reflects biological properties of LSCs that confer resistance to standard AML therapy. The LSC17 NanoString assay will allow rapid risk assessment at diagnosis, enabling recommendation of more intensified investigational therapies to be directed to high-score patients predicted to have resistant disease, while sparing low-score patients unnecessary added toxicity. Furthermore, our analysis of the ALFA-0701 trial data demonstrates the utility of the LSC17 score as a tool for patient selection, and can probably be extended to additional patient cohorts treated with other experimental therapies as data becomes available in future studies. Finally, the LSC3 score will allow evaluation of the possible benefits of post-remission therapy or aSCT in CR1 for high-risk CN-LMR patients, or could be used in upfront therapy decisions at centres where CN-LMR patients can be rapidly identified by molecular diagnostics. Overall, incorporation of the LSC17 and LSC3 scores into risk determination algorithms for newly diagnosed AML patients will facilitate the development and clinical testing of novel anti-leukaemia therapies in the ongoing effort to prevent relapse and increase cure rates.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 September; accepted 2 November 2016.

Published online 7 December 2016.

- Ferrara, F. & Schiffer, C. A. Acute myeloid leukaemia in adults. *Lancet* **381**, 484–495 (2013).
- Grimwade, D. *et al.* Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. *Blood* **116**, 354–365 (2010).
- Döhner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453–474 (2010).
- Röllig, C. *et al.* Long-term prognosis of acute myeloid leukemia according to the new genetic risk classification of the European LeukemiaNet recommendations: evaluation of the proposed reporting system. *J. Clin. Oncol.* **29**, 2758–2765 (2011).
- Walter, R. B. *et al.* Resistance prediction in AML: analysis of 4601 patients from MRC/NCR, HOVON/SAKK, SWOG and MD Anderson Cancer Center. *Leukemia* **29**, 312–320 (2015).
- Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).
- Saito, Y. *et al.* Identification of therapeutic targets for quiescent, chemotherapy-resistant human leukemia stem cells. *Sci. Transl. Med.* **2**, 17ra9 (2010).
- Li, L. *et al.* SIRT1 activation by a c-MYC oncogenic network promotes the maintenance and drug resistance of human FLT3-ITD acute myeloid leukemia stem cells. *Cell Stem Cell* **15**, 431–446 (2014).

9. Fong, C. Y. *et al.* BET inhibitor resistance emerges from leukaemia stem cells. *Nature* **525**, 538–542 (2015).
10. Lechman, E. R. *et al.* miR-126 regulates distinct self-renewal outcomes in normal and malignant hematopoietic stem cells. *Cancer Cell* **29**, 214–228 (2016).
11. Eppert, K. *et al.* Stem cell gene expression programs influence clinical outcome in human leukemia. *Nature Med.* **17**, 1086–1093 (2011).
12. Sarry, J. E. *et al.* Human acute myelogenous leukemia stem cells are rare and heterogeneous when assayed in NOD/SCID/IL2R γ -deficient mice. *J. Clin. Invest.* **121**, 384–395 (2011).
13. Laurenti, E. *et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nature Immunol.* **14**, 756–763 (2013).
14. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
15. Verhaak, R. G. *et al.* Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica* **94**, 131–134 (2009).
16. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
17. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
18. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
19. Metzeler, K. H. *et al.* An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* **112**, 4193–4201 (2008).
20. Grimwade, D., Ivey, A. & Huntly, B. J. Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance. *Blood* **127**, 29–41 (2016).
21. Papaemmanuil, E. *et al.* Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
22. Levine, J. H. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
23. Gentles, A. J., Plevritis, S. K., Majeti, R. & Alizadeh, A. A. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *J. Am. Med. Assoc.* **304**, 2706–2715 (2010).
24. Jung, N., Dai, B., Gentles, A. J., Majeti, R. & Feinberg, A. P. An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. *Nature Commun.* **6**, 8489 (2015).
25. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnol.* **26**, 317–325 (2008).
26. Cornelissen, J. J. *et al.* The European LeukemiaNet AML Working Party consensus statement on allogeneic HSCT for patients with AML in remission: an integrated-risk adapted approach. *Nature Rev. Clin. Oncol.* **9**, 579–590 (2012).
27. Kohlmann, A. *et al.* Gene expression profiling in AML with normal karyotype can predict mutations for molecular markers and allows novel insights into perturbed biological pathways. *Leukemia* **24**, 1216–1220 (2010).
28. Castaigne, S. *et al.* Effect of gemtuzumab ozogamicin on survival of adult patients with *de-novo* acute myeloid leukaemia (ALFA-0701): a randomised, open-label, phase 3 study. *Lancet* **379**, 1508–1516 (2012).
29. Hills, R. K. *et al.* Addition of gemtuzumab ozogamicin to induction chemotherapy in adult patients with acute myeloid leukaemia: a meta-analysis of individual patient data from randomised controlled trials. *Lancet Oncol.* **15**, 986–996 (2014).
30. Kilo, J. M. *et al.* Association between mutation clearance after induction therapy and outcomes in acute myeloid leukemia. *J. Am. Med. Assoc.* **314**, 811–822 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the Ontario Institute for Cancer Research with funds from the province of Ontario, the Cancer Stem Cell Consortium with funding from the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-047), and the Canadian Institutes of Health Research (CSC-105367), Canadian Cancer Society, Terry Fox Foundation, a Canada Research Chair to J.E.D., the Philip S. Orsino Chair in Leukemia Research to M.D.M., and a Collaborative Translational Cancer Research Grant from the Princess Margaret Cancer Centre (formerly Ontario Cancer Institute). This research was funded in part by the Leukemia & Lymphoma Society of Canada (493946) and the Stem Cell Network (492019), Ontario Graduate Scholarships, and the Ontario Ministry of Health and Long Term Care (OMOHLTC). The views expressed do not necessarily reflect those of the OMOHLTC. L.B. was supported in part by the Deutsche Forschungsgemeinschaft (Heisenberg-Professur BU 1339/8-1). T.H. was supported by the Wilhelm-Sander-Stiftung (grant 2013.086.1). K.M. and W.H. received grant support from Deutsche Forschungsgemeinschaft (DFG SFB 1243). We thank The Centre for Applied Genomics (Hospital for Sick Children) and the Princess Margaret Genomics Centre for the generation of GE data for the PM sorted cell fractions and validation cohort. We thank M. Pintilie for discussions regarding time-dependent covariates in survival analysis. We thank S. Geffroy for technical support and running the microarrays for the ALFA-0701 trial cohort.

Author Contributions S.W.K.N. developed the signature derivation workflow, identified, refined and validated prognostic and predictive signatures, designed the custom NanoString assay, processed and analysed GE data, and performed statistical analyses and bioinformatics. A.M., W.C.C., J.M. and A.P. carried out functional xenograft transplantation, RNA extraction for GE analysis, and provided technical support for experiments. J.A.K., N.I., A.A., V.G., A.D.S., A.C.S., K.W.Y. and M.D.M. provided clinical annotations for the PM AML cohort. M.D.M. provided PM AML samples. S.W.K.N., J.C.Y.W., J.E.D. and M.D.M. interpreted the data. W.H., W.E.B., B.W., T.B., D.G., L.B., K.M., T.H. and C.B. provided clinical annotations for the GSE15434 and GSE12417 data sets. M.C., C.P. and H.D. provided GE and clinical data for the ALFA-0701 trial cohort. P.J.M.V. and B.L. provided clinical annotations for the GSE6891 data set. J.C.Y.W. and J.E.D. supervised the study. S.W.K.N. and J.C.Y.W. wrote the paper. A.M., J.A.K., P.W.Z., J.E.D. and M.D.M. revised the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.C.Y.W. (jwang@uhnresearch.ca).

Reviewer Information *Nature* thanks F. Holstege, G. Schuurhuis and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Patient samples. All biological samples were collected with informed consent according to procedures approved by the Research Ethics Board of the University Health Network (UHN; REB# 01-0573-C) and viably frozen in the PM Leukaemia Bank. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment. **Xenotransplantation assays.** Eighty-three clinical samples (81 peripheral blood (PB), 1 bone marrow (BM), 1 peritoneal fluid) obtained from 78 patients were stained with the following antibodies (all from BD, dilution 1:100, catalogue number in parentheses): anti-CD3-FITC (349201), anti-CD34-APC (340441), anti-CD38-PE (347687). Each sample was sorted on a FACSaria III (BD Biosciences) into four fractions based on CD34/CD38 expression; a total of 227 fractions with sufficient cell numbers were tested in xenotransplantation assays. The clinical characteristics of these patients are provided in Extended Data Fig. 1a. Sixty-two samples were diagnostic, 16 were obtained following relapse and 5 after unsuccessful induction treatment. Paired diagnosis–relapse samples were included from 3 patients. Diagnosis samples from BM and PB of 1 patient were included. Two PB relapse samples collected at different times from 1 patient were included.

Animal experiments were performed in accordance with institutional guidelines approved by the UHN Animal Care Committee. Eight to 12-week-old female NSG mice were sublethally irradiated (225 cGy) 24 h before intrafemoral injection of sorted AML cell fractions. Mice were killed 12 weeks post-transplant and human cell engraftment in the injected right femur was assessed by flow cytometry using human-specific antibodies (all used at 1:100 dilution, all from BD unless stated otherwise, catalogue number in parentheses): anti-CD3-FITC (349201), anti-CD19-PE (349209), anti-CD33-PE-Cy5 (Beckman Coulter PNIM2647U), anti-CD45-APC (340943), anti-CD38-PE-Cy7 (1:200, 335790) and anti-CD34-APC-Cy7 (custom made by BD). AML grafts were defined as $\geq 0.1\%$ human CD45⁺CD3⁺ cells, with $\geq 90\%$ CD33 expression. Sorted fractions were defined as LSC⁺ if transplanted cells generated an AML graft in 1 or more mice; the remaining fractions were defined as LSC[−]. All flow cytometric analysis was performed on a BD LSR II.

GE profiling of cell fractions. RNA was extracted using Qiagen RNeasy mini kits (catalogue 74106) and was subjected to GE analysis using Illumina HumanHT-12 v4 microarrays to investigate ~47,000 targets corresponding to ~30,000 genes. The resultant fluorescence intensity profiles were subjected to variance stabilization and robust spline normalization using the lumi 2.16.0 R package³¹. All data was put into the log base-2 scale. Differential GE analysis was performed using the limma 3.20.9 package³² in R. Specifically, Smyth's moderated *t*-test was used with Benjamini–Hochberg multiple testing correction to compare GE profiles of LSC⁺ versus LSC[−] fractions. Relative proportions of GE programs of stem/progenitor and mature cell types purified from human umbilical cord blood (GEO accessions GSE42414 (ref. 13) and GSE24759 (ref. 14)) composing AML GE profiles were assessed using the Perturbation model³³.

Signature training. For signature development, we used published GE profiles of diagnostic samples obtained from 537 patients with *de novo* AML treated with curative intent (GSE6891). Clinical annotations for 521 cases were provided by the authors¹⁵. Of these, we removed 23 cases of myelodysplastic syndrome refractory anaemia with excess blasts (MDS-RAEB), 2 cases due to missing WBC count data, and 1 because there was no raw GE data available for download, leaving 495 cases for analysis (Extended Data Fig. 1f). The GE data from this study were generated using Affymetrix Human Genome (HG) U133 Plus 2.0 GeneChips. The probes available on this array capture 89 of the 104 LSC associated genes (43 of the 48 enriched in LSC⁺ cell fractions) (Extended Data Table 1). Raw Affymetrix CEL files were imported using the affy 1.42.3 R package³⁴ and processed with the gcrma 2.36.0 package³⁵ in R, using version 17 of the custom chip definition files (CDF) for the HG-U133 Plus 2.0 platform from the University of Michigan³⁶.

For each gene, the probe set with the highest average GE in the training data was selected to represent that gene. To extract a core subset of genes from among the 43 that were more highly expressed in LSC⁺ cell fractions that best explained patient outcomes in the training cohort, we used a linear regression technique based on the LASSO algorithm as implemented in the glmnet 1.9-8 R package^{16,17}, while enabling leave-one-out cross-validation to fit a Cox regression model. A minimal subset of 17 genes was selected whose weighted combined GE (LSC17 score) was highly correlated to survival outcomes in the training cohort.

The LSC17 score is calculated for each patient as a linear combination of GE of these 17 genes weighted by regression coefficients that were estimated from the training data as follows: LSC17 score = $(DNMT3B \times 0.0874) + (ZBTB46 \times -0.0347) + (NYNRIN \times 0.00865) + (ARHGAP22 \times -0.0138) + (LAPTM4B \times 0.00582) + (MMRN1 \times 0.0258) + (DPYSL3 \times 0.0284) + (KIAA0125 \times 0.0196) + (CDK6 \times -0.0704) + (CPXM1 \times -0.0258) + (SOSC2 \times 0.0271) + (SMIM24 \times -0.0226) + (EMP1 \times 0.0146) + (NGFRAP1 \times 0.0465) + (CD34 \times 0.0338) + (AKRIC3 \times -0.0402) + (GPR56 \times 0.0501)$. As above- and below-median scores in the training

cohort were associated with adverse and favourable cytogenetic risk, respectively, a median threshold was used to discretize scores into high and low groups.

An optimized sub-signature was identified by applying the above described regression procedure to CN-LMR cases from GSE6891 with OS >30 days ($n = 44$), while restricting the analysis to the LSC17 genes. A new equation resulted for computing CN-LMR patient-specific risk scores: LSC3 score = $(DPYSL3 \times 0.3) + (AKRIC3 \times -0.0477) + (NYNRIN \times 0.194)$.

Similarly, a retrained treatment response score comprising 6 of the LSC17 genes was derived by applying the above described regression workflow to a randomly chosen half of the PM AML cohort: initial induction response score = $-6.58 + (MMRN1 \times 0.0442) + (KIAA0125 \times 0.0814) + (CD34 \times 0.104) + (GPR56 \times 0.208) + (LAPTM4B \times 0.168) + (NYNRIN \times 0.121)$.

Signature testing: microarray data processing and analysis. The LSC17 score was initially validated against three published clinically annotated AML cohorts with available microarray GE data (one from TCGA¹⁸ and two from GSE12417 (ref. 19)), while the LSC3 score was tested on an independent CN-LMR subset from GSE15434 (ref. 27). Treatment protocols and the criteria used for cytogenetic/molecular risk classification for each cohort have been previously described^{18,19,27}. Raw Affymetrix CEL files (generated on the HG-U133 Plus 2.0 array) containing GE data of a cohort of *de novo* AML patients of all cytogenetic risk groups¹⁸ along with clinical data were downloaded from the TCGA AML data portal ($n = 183$; Extended Data Table 2). Raw Affymetrix CEL files (generated on the HG-U133 A, B, and Plus 2.0 arrays) containing GE data for two independent cohorts of CN-AML cases¹⁹ were downloaded (GSE12417) and clinical annotations were provided by the authors. Of the 163 GE profiles in GSE12417 CN-AML cohort 1, we removed 2 PB, 1 MDS-RAEB, and 4 other cases with missing clinical data, leaving 156 for analysis (Extended Data Table 3a). For this cohort, the GE data generated on the HG-U133 A and B arrays were merged. The same inclusion criteria for analysis were applied to the 79-patient GSE12417 CN-AML cohort 2, leading to the removal of 1 MDS, 5 PB, and 3 other cases due to missing clinical data, leaving 70 for analysis (Extended Data Table 3b). A data set of 70 CN-LMR HG-U133 Plus 2.0 array profiles was downloaded (GSE15434), with clinical data provided by the authors²⁷. In addition, clinical and HG-U133 Plus 2.0 microarray GE data for AML patients treated in the ALFA-0701 trial were provided by the authors²⁸ (Extended Data Table 8a). All microarray data were normalized as described for the training data set (GSE6891). Signature scores (LSC17 or LSC3) were calculated for each patient in the validation cohorts using the linear equations derived during signature training and a median threshold.

Signature testing: RNA-seq data processing and analysis. One-hundred and sixty-nine patients in the TCGA AML cohort had both microarray and RNA-seq GE data available. The Illumina GA-IIX RNA-seq profiles normalized to reads per kilobase of transcript per million mapped reads (RPKM) were downloaded from the TCGA AML data portal. A value of 1 was added to the RPKM values before applying a log-transformation to the base-2 scale. For each gene, the entry with the maximum mean GE in the data set was used for computing LSC17 scores.

NanoString assay design and GE profiling. We submitted the 17 Affymetrix probe set identifiers associated with the LSC17 score (Extended Data Table 1), along with reference genes chosen to cover a wide range of expression levels in AML³⁷, to NanoString Technologies²⁵ for custom codeset creation. The 100 base pair (bp) NanoString probes were fabricated to overlap or be proximal to the corresponding Affymetrix probe target regions. On each 12-lane NanoString cartridge implementing this codeset design, a single lane was reserved for a control composed of an equal parts mixture of 26 synthetic 100 bp DNA oligonucleotides designed to resemble the target transcripts (Integrated DNA Technologies, 1.8 pM per oligonucleotide), against which the GE across all cartridges was normalized to minimize inter-cartridge variability as done by others^{38,39}. For each of the remaining 11 lanes, 100 ng, 150 ng, or 250 ng of RNA per sample (5 μ l) was incubated with 20 μ l of reporter probe and 5 μ l of capture probe mix (supplied by the manufacturer) at 65 °C for 16 to 24 h for hybridization on the nCounter Prep Station (version 4.0.11.1). After hybridization, excess probes were washed out using a 2-step magnetic bead-based purification strategy according to the manufacturer's protocol, and purified target/probe complexes were immobilized on the NanoString cartridge for data collection. Transcript counts were determined using the nCounter Digital Analyzer (version 2.1.2.3) at the high-resolution setting. Specifically, digital images were processed with final barcode counts tabulated in reporter code count (RCC) output files.

Signature testing: NanoString data processing and analysis. The NanoString assay was performed using RNA from bulk mononuclear cells obtained from 307 banked diagnostic samples collected from patients treated at PM with curative intent between 1999 and 2012 (Extended Data Table 7a). Patients were excluded if they received any cytoreductive treatment other than hydroxyurea or died within one month of starting therapy. RCC files containing raw transcript counts from each cartridge were analysed using the nSolver analysis software (version 2.0.72)

for quality control (QC) and normalization purposes using default settings for GE analysis. Specifically, RCC files for each cartridge along with a reporter library file containing codeset probe annotations were imported into nSolver. The software was used to normalize the captured transcript counts to the geometric mean of the reference genes included in our assay and the codeset's internal positive controls, and to check for imaging, binding, positive spike-in, and normalization quality. The control lane of each cartridge was processed in the same manner as the RNA lanes using the nSolver software without normalization to reference genes.

The output files from nSolver were read into R for further QC, normalization, and data processing. An RNA input correction step was used to adjust the GE counts of each cartridge to the reference amount of 100 ng RNA. The control lanes for cartridges 1 to 3 were used as blank lanes to estimate per-probe background noise. None of the signature or reference probes exhibited high background counts (that is, <3 standard deviations (s.d.) above the geometric mean of the codeset's 8 internal negative control probes) and thus no background subtraction was required. In lanes where RNA was present, all signature and reference probe counts were well above 3 s.d. over background. The coefficient of variation (CV; s.d. divided by mean GE) and maximum fold change (MFC; maximum divided by minimum GE) were used to quantify GE variation. All reference probes had lower CV and MFC values compared to signature probes and most codeset controls while spanning a sufficiently large range of signature probe GE.

To batch-correct control oligonucleotide counts, multiplicative corrective constants were computed and applied to each batch of control lanes according to the oligonucleotide preparation schedule. Specifically, the oligonucleotide counts of each batch of control lanes were scaled by a ratio of geometric means between the oligonucleotide counts in each batch and that of all control lanes. We next used the batch-corrected control lanes to minimize inter-cartridge technical variation in RNA counts. The geometric mean of the corrected oligonucleotide counts in the control lane of cartridge 5 (arbitrarily chosen) was divided by the same summary value corresponding to each of the other cartridges to produce per-cartridge scaling factors. The RNA and oligonucleotide counts of each cartridge were then adjusted using these factors by means of multiplication, thereby minimizing batch induced GE variation. A final round of normalization to the reference genes was then performed by adjusting the GE counts in all 307 RNA lanes in the data set using a ratio of geometric means between the reference GE counts in each cartridge and that of all cartridges. The fully normalized GE counts were log₂-transformed after incrementing by 1. Signature scores (LSC17 or LSC3) were computed for each patient using the scaled data.

PM cohort treatment details. All patients received induction chemotherapy with a 3+7 backbone (daunorubicin 60 mg/m² intravenously (i.v.) daily × 3 d + cytarabine (ara-C) 200 mg/m² i.v. daily × 7 d. A minority of patients were enrolled in clinical trials employing 3+7 with GO (*n* = 7) or midostaurin (*n* = 3). Barring contraindications, patients achieving CR went on to receive two cycles of consolidation chemotherapy with daunorubicin 45 mg/m² i.v. on days 1–2 + ara-C 3 g/m² every 12 h on days 1, 3, 5. Patients with core binding factor leukaemia received one cycle of this consolidation followed by two cycles of ara-C (3 g/m² every 12 h on days 1, 3, 5). For APL patients, induction and the first consolidation cycle included all-*trans* retinoic acid (ATRA) 45 mg/m² daily × 28 d, daunorubicin 60 mg/m² i.v. daily × 3 d and ara-C 100 mg/m² i.v. daily × 7 d. The second consolidation cycle included ATRA × 28 d, daunorubicin 45 mg/m² i.v. daily on days 1–3 and ara-C 1.5 g/m² every 12 h on days 1, 3, 5. For patients ≥ 60 years of age with WBC count < 10, ara-C was omitted from induction and consolidation. For APL patients with initial WBC count < 10, maintenance therapy consisted of ATRA 45 mg/m²/d × 7 d on alternating weeks × 9 months. For all others, maintenance involved 21 monthly cycles of 6-mercaptopurine 75 mg/m²/d daily for 21 d and methotrexate 20 mg/m²/d once weekly; every other cycle included ATRA 45 mg/m²/d × 14 d. aSCT was performed in CR1 for high-risk patients, typically those with secondary AML, adverse cytogenetics, or normal karyotype with poor prognostic molecular features.

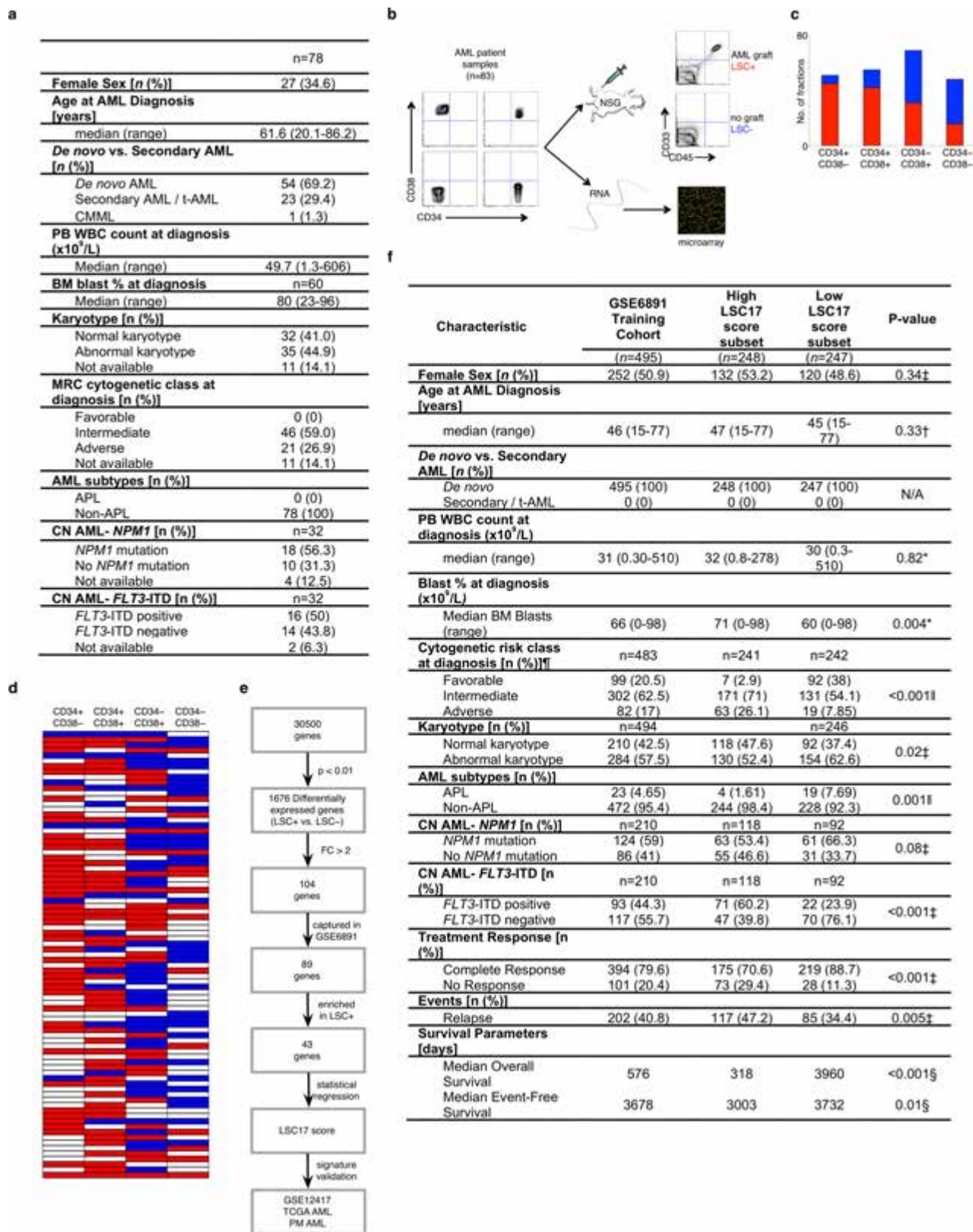
Statistical analysis. All statistical analyses were performed in R 3.1.0 (ref. 40). The Spearman rank method of correlation was used unless specified otherwise. Various two-tailed tests were used to evaluate the differences in baseline clinical characteristics between patients with high versus low LSC17 scores. OS was defined as the time from AML diagnosis until death from any cause or last clinical follow-up. EFS was defined as the time from AML diagnosis until an event (that is, induction failure, relapse or death from any cause) or last follow-up. RFS was defined as the time from CR1 until relapse or death (regardless of cause) or last clinical follow-up⁴¹. Univariate survival analysis was performed using the Kaplan–Meier and CPH models and a median threshold, with comparisons performed using Mantel–Cox log-rank tests. For multivariate analyses, covariates for CPH models included LSC17 or LSC3 score, as well as established risk factors (for example, age and WBC count at diagnosis, *de novo* versus secondary AML, cytogenetic risk group, and *NPM1* and *FLT3*-ITD mutational status). The intermediate

risk subgroup was used as a reference against which other risk subgroups were compared, unless specified otherwise. Wald's test was used to evaluate the significance of HRs, while violation of the proportional hazards assumption was detected by examining Schoenfeld residuals, and eliminated by setting offending parameters as stratifying variables in the model as done by others^{18,19}. Cumulative incidence analysis of relapse and death as competing risks was assessed using Gray (univariate) and Fine–Gray (multivariate) methods^{42,43}, as implemented in the *cmprsk*⁴⁴ and *riskRegression*⁴⁵ R packages in the EZR software⁴⁶. The impact of aSCT on OS, EFS and RFS, was assessed by encoding transplant as a time-dependent covariate in uni- and multi-variate Mantel–Byar⁴⁷ and Andersen–Gill⁴⁸ models, respectively, where univariate results were visualized using Simon–Makuch plots⁴⁹. All survival analyses were performed using the survival 2.38-1 R package⁵⁰. In comparing the LSC17 score to other reported signatures, custom CDFs were used to summarize microarray probe expression for each gene, unless specified otherwise.

In analyses assessing prediction of treatment response, uni- and multi-variate logistic regression models were used with the bootstrap-adjusted AUROC metric to determine the ability of various parameters to predict initial induction response. The rms 4.4-1 R package⁵¹ was used for logistic regression analysis, while the *proC* 1.8 and *PredictABEL* 1.2-2 R packages were used for ROC curve analyses^{52,53}. Relative importance of individual covariates in multivariate logistic regression models was estimated by examining the partial Wald Chi-squared statistic as done by others⁵.

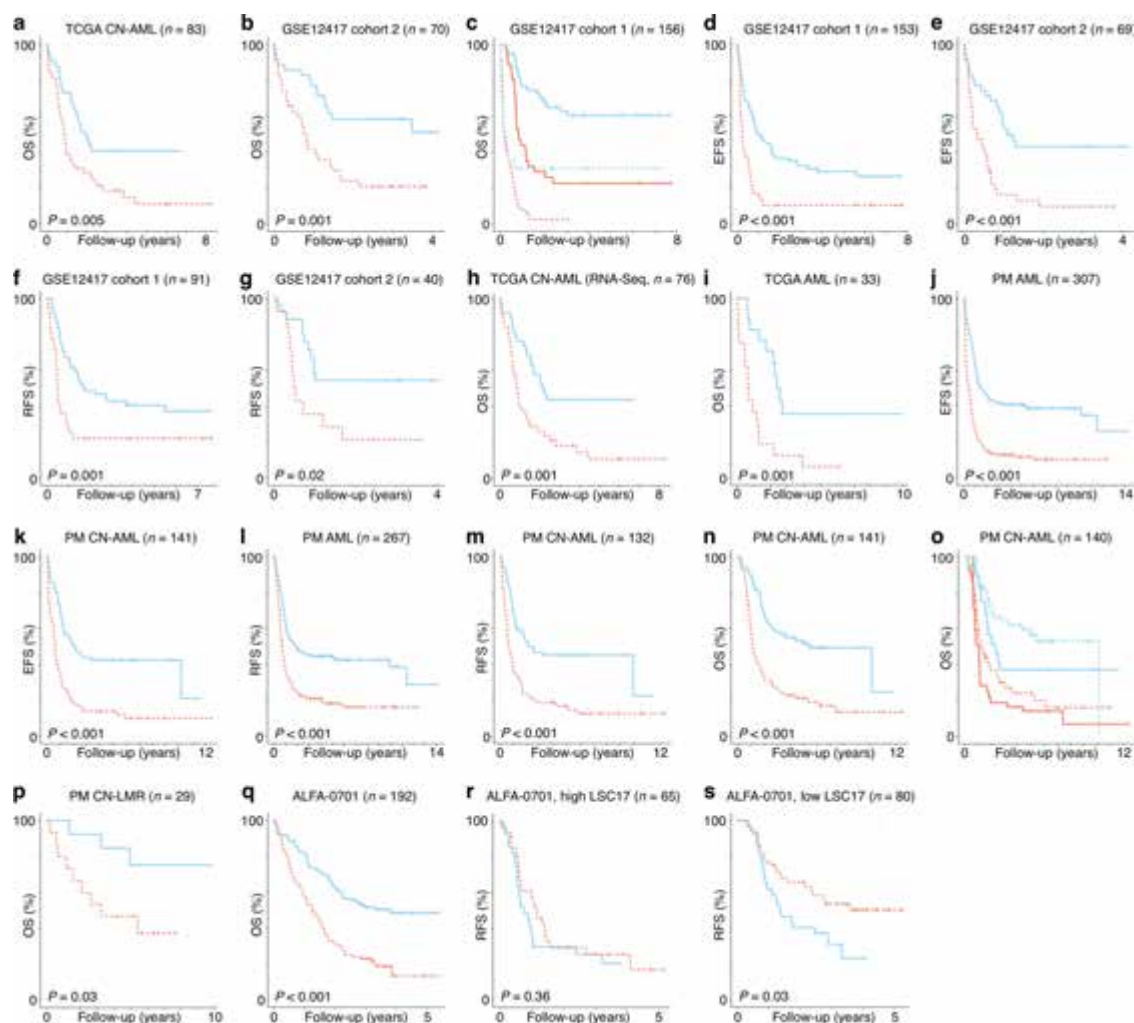
Data availability. All raw and normalized GE data that support the findings of this study have been deposited in the GEO SuperSeries under accession number GSE76009 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76009>).

- Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
- Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
- Qiao, W. *et al.* PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLOS Comput. Biol.* **8**, e1002838 (2012).
- Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
- Wu, J., Irizarry, R., MacDonald, J. & Gentry, J. Gcrma: background adjustment using sequence information. R package version 2.36.0 (2016).
- Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).
- Macrae, T. *et al.* RNA-seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells. *PLoS ONE* **8**, e72884 (2013).
- Scott, D. W. *et al.* Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood* **123**, 1214–1217 (2014).
- Nielsen, T. *et al.* Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC Cancer* **14**, 177 (2014).
- R Development Core Team. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
- Cheson, B. D. *et al.* Revised recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia. *J. Clin. Oncol.* **21**, 4642–4649 (2003).
- Gray, R. J. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Stat.* **16**, 1141–1154 (1988).
- Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999).
- Gray, B. *cmprsk*: subdistribution analysis of competing risks. R package version 2.2-7 (2014).
- Gerds, T. A. & Scheike, T. H. *riskRegression*: risk regression for survival analysis. R package version 0.0.8 (2016).
- Kanda, Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transplant.* **48**, 452–458 (2013).
- Mantel, N. & Byar, D. Evaluation of response-time data involving transient states: an illustration using heart transplant data. *J. Am. Stat. Assoc.* **69**, 81–86 (1974).
- Andersen, P. & Gill, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10**, 1100–1120 (1982).
- Simon, R. & Makuch, R. W. A non-parametric graphical representation of the relationship between survival and the occurrence of an event: application to responder versus non-responder bias. *Stat. Med.* **3**, 35–44 (1984).
- Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model* (Springer, 2000).
- Harrell, F. E. Jr. *rms*: regression modeling strategies. R package version 4.4-1 (2016).
- Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
- Kundu, S., Aulchenko, Y. S., van Duijn, C. M. & Janssens, A. C. PredictABEL: an R package for the assessment of risk prediction models. *Eur. J. Epidemiol.* **26**, 261–264 (2011).



Extended Data Figure 1 | Overview of LSC signature training and testing. **a**, Clinical characteristics of the 78 patients analysed by xenotransplantation and microarray GE analysis. CMML, chronic myelomonocytic leukaemia; t-AML, therapy-associated AML; CN, cytogenetically normal. **b**, Schematic of the experimental protocol. **c**, **d**, Summary of functionally defined LSC⁺ and LSC⁻ fractions in each phenotypic cell population as a whole (**c**) and for each patient (**d**). Red and blue denote LSC⁺ and LSC⁻, respectively. In **d**, each row represents fractions sorted from one patient sample. White boxes denote fractions

that were not included in the analysis due to insufficient cell numbers for xenotransplantation and/or insufficient RNA. **e**, Strategy used to identify and test the 17 LSC signature genes. **f**, Key clinical characteristics of the GSE6891 signature training cohort. *P value calculated using the Wilcoxon rank-sum test; †P value calculated using the Student's *t*-test; ‡P value calculated using Pearson's chi-squared test; §P value calculated using log-rank test; ||P value calculated using Fisher's exact test; ¶cytogenetic risk groups were defined as per GSE6891 investigators¹⁵.



Extended Data Figure 2 | LSC17 and LSC3 scores are associated with survival in multiple AML cohorts. a–n, q, Kaplan–Meier estimates of OS, EFS or RFS according to LSC17 scores in various patient cohorts, as indicated. In **c**, patients were also analysed according to whether or not CR was achieved after initial treatment (no CR, dotted lines; CR, solid lines). **i**, The subset of patients in the TCGA AML cohort with no clear genomic classification as defined previously²¹. **o**, Simon and Makuch estimates of

OS, according to LSC17 scores and whether or not patients received aSCT (no aSCT, dotted lines; aSCT, solid lines). **p**, Kaplan–Meier estimates of OS of CN-LMR patients, according to LSC3 scores. In **a–q**, patients with scores above and below the median in each cohort are shown by red and blue lines, respectively. **r, s**, Kaplan–Meier estimates of RFS for patients with high (**r**) or low (**s**) LSC17 scores treated with standard chemotherapy with (red lines) or without (blue lines) addition of GO.

Extended Data Table 1 | List of 104 DE LSC genes

Gene Symbol	Entrez ID	Illumina Probe ID*	Log ₂ Fold Change†	P-value‡	Affymetrix Probeset IDs	Signature Gene
CD34	947	ILMN_1732799	2.15	<0.0001	209543 s at	LSC17
SPINK2	6691	ILMN_1763516	1.99	<0.0001	206310 at	N/A
LAPTM4B	55353	ILMN_2101832	1.8	<0.0001	214039 s at	LSC17
HOKA5	3202	ILMN_1753613	1.72	<0.0001	213844 at	N/A
GUCY1A3	2982	ILMN_1808590	1.62	<0.0001	229530 at	N/A
SHANK3	85358	ILMN_2317581	1.59	<0.0001	227923 at	N/A
ANGPT1	284	ILMN_1677723	1.51	<0.0001	205609 at	N/A
ARHGAP22	58504	ILMN_1676361	1.48	<0.0001	206298 at	LSC17
LOC284422	284422	ILMN_1774375	1.45	<0.0001	231982 at	LSC17
MYCN	4613	ILMN_2219767	1.41	<0.0001	209757 s at	N/A
MAMDC2	256691	ILMN_1679391	1.4	<0.0001	228885 at	N/A
PRSSL1	400668	ILMN_1673605	1.4	<0.0001	N/A	N/A
KIAA0125	9834	ILMN_1707491	1.4	<0.0001	206478 at	LSC17
GPSM1	26086	ILMN_1709307	1.38	<0.0001	226043 at	N/A
HOKA9	3205	ILMN_1739582	1.38	<0.0001	N/A	N/A
MMRN1	22915	ILMN_1660114	1.36	<0.0001	205612 at	LSC17
FSCN1	6624	ILMN_1808707	1.32	<0.0001	210933 s at	N/A
DNMT3B	1789	ILMN_2328972	1.31	<0.0001	220668 s at	LSC17
HOKA6	3203	ILMN_1815570	1.28	<0.0001	208557 at	N/A
AIF1L	83543	ILMN_3246401	1.25	<0.0001	223075 s at	N/A
SOC5	8835	ILMN_1798926	1.24	<0.0001	203373 at	LSC17
CDK6	1021	ILMN_1802615	1.23	<0.0001	224851 at	LSC17
FAM50B	138311	ILMN_1757440	1.2	<0.0001	229002 at	N/A
NGFRAP1	27018	ILMN_2370091	1.2	<0.0001	217963 s at	LSC17
C3orf54	389119	ILMN_1690454	1.2	<0.0001	229507 at	N/A
CPXM1	56265	ILMN_1712046	1.2	<0.0001	227860 at	LSC17
TNFRSF4	7293	ILMN_2112256	1.2	<0.0001	214228 x at	N/A
ZBTB46	140685	ILMN_1710092	1.19	<0.0001	227329 at	LSC17
DPYSL3	1809	ILMN_1679262	1.18	<0.0001	201431 s at	LSC17 & LSC3
NYNRIN	57523	ILMN_3236858	1.15	<0.0001	220911 s at	LSC17 & LSC3
COL24A1	255631	ILMN_1810996	1.13	<0.0001	238732 at	N/A
FAM30A	29064	ILMN_3187535	1.11	<0.0001	N/A	N/A
C10orf140	387640	ILMN_3238861	1.1	<0.0001	N/A	N/A
SPNS2	124976	ILMN_3301749	1.07	<0.0001	225671 at	N/A
GPR56	9289	ILMN_2384122	1.07	0.00054	212070 at	LSC17
AKR1C3	8644	ILMN_1713124	1.06	<0.0001	209160 at	LSC17 & LSC3
FLT3	2322	ILMN_1766363	1.05	<0.0001	206674 at	N/A
TFPI	7035	ILMN_1707124	1.05	<0.0001	213258 at	N/A
KCNK17	89822	ILMN_1717702	1.04	<0.0001	224049 at	N/A
EPDR1	54749	ILMN_1675797	1.03	<0.0001	223253 at	N/A
C1orf150	148823	ILMN_1762204	1.02	<0.0001	N/A	N/A
BIVM	54841	ILMN_2214098	1.02	<0.0001	222761 at	N/A
H2AFY2	55506	ILMN_1705570	1.02	<0.0001	218445 at	N/A
VWF	7450	ILMN_1752755	1.02	0.000103	202112 at	N/A
EMP1	2012	ILMN_1801616	1.01	<0.0001	201324 at	LSC17
RAGE	5891	ILMN_1745282	1.01	<0.0001	205130 at	N/A
ATP8B4	79695	ILMN_1783956	1.01	<0.0001	220416 at	N/A
GATA2	2624	ILMN_2102670	1	<0.0001	209710 at	N/A
SLC25A37	51312	ILMN_1715969	-1.01	<0.0001	222528 s at	N/A
SGK	6446	ILMN_3305938	-1.01	<0.0001	201739 at	N/A
LOC652694	652694	ILMN_1680274	-1.01	<0.0001	N/A	N/A
ITPR3	3710	ILMN_1815500	-1.02	<0.0001	201187 s at	N/A
LOC654103	654103	ILMN_1802808	-1.02	<0.0001	N/A	N/A
CXCR4	7852	ILMN_1801584	-1.04	<0.0001	217028 at	N/A
FCRL3	115352	ILMN_1691693	-1.05	<0.0001	N/A	N/A
RBM38	55544	ILMN_2404049	-1.05	<0.0001	212430 at	N/A
LILRA5	353514	ILMN_2357419	-1.06	<0.0001	215838 at	N/A
IL18RAP	8807	ILMN_1721762	-1.06	<0.0001	207072 at	N/A
CDC109B	55013	ILMN_1801766	-1.08	<0.0001	218802 at	N/A
ISG20	3669	ILMN_1659913	-1.09	<0.0001	33304 at	N/A
MTSS1	9788	ILMN_2073289	-1.09	<0.0001	203037 s at	N/A
CECR1	51816	ILMN_1751851	-1.1	<0.0001	219505 at	N/A
ADAM19	8728	ILMN_1713751	-1.1	<0.0001	209765 at	N/A
FCGR2A	2212	ILMN_1666932	-1.11	<0.0001	N/A	N/A
AIM2	9447	ILMN_1681301	-1.11	<0.0001	206513 at	N/A
NPL	80896	ILMN_1782070	-1.14	<0.0001	223405 at	N/A
IL10RA	3587	ILMN_1652825	-1.15	<0.0001	204912 at	N/A
CTSL1	1514	ILMN_1812995	-1.16	<0.0001	202087 s at	N/A
GNLV	10578	ILMN_1708779	-1.19	<0.0001	205495 s at	N/A
CKAP4	10970	ILMN_1790891	-1.19	<0.0001	200999 s at	N/A
ADM	153	ILMN_1706934	-1.19	<0.0001	202912 at	N/A
KLRB1	3620	ILMN_2079655	-1.19	<0.0001	214470 at	N/A
SLC15A3	51296	ILMN_2085862	-1.21	<0.0001	219593 at	N/A
FGR	2268	ILMN_1795158	-1.22	<0.0001	208438 s at	N/A
FCRLA	84824	ILMN_1691071	-1.22	<0.0001	235372 at	N/A
IL2RB	3560	ILMN_1684349	-1.23	<0.0001	205291 at	N/A
CXCL16	58191	ILMN_1728478	-1.24	<0.0001	223454 at	N/A
SLC4A1	6521	ILMN_1772809	-1.24	<0.0001	205592 at	N/A
GZMH	2999	ILMN_1731233	-1.27	<0.0001	210321 at	N/A
FLJ22662	79687	ILMN_1707286	-1.27	<0.0001	218454 at	N/A
LOC647506	647506	ILMN_3240375	-1.28	<0.0001	N/A	N/A
GIMAP4	55303	ILMN_1748473	-1.29	<0.0001	219243 at	N/A
JAZF1	221895	ILMN_1682727	-1.32	<0.0001	225798 at	N/A
CTSH	1512	ILMN_2390853	-1.33	<0.0001	202295 s at	N/A
GZMA	3001	ILMN_1779324	-1.35	<0.0001	205488 at	N/A
CHST15	51363	ILMN_1670926	-1.35	<0.0001	203066 at	N/A
AQP9	366	ILMN_1715068	-1.4	<0.0001	205568 at	N/A
CD247	919	ILMN_1676924	-1.41	<0.0001	210031 at	N/A
BCL6	604	ILMN_1737314	-1.42	<0.0001	203140 at	N/A
SLC7A7	9056	ILMN_1810275	-1.43	<0.0001	204588 s at	N/A
E2F2	1870	ILMN_1777233	-1.45	<0.0001	228361 at	N/A
LOC647450	647450	ILMN_1699214	-1.45	<0.0001	N/A	N/A
GZMB	3002	ILMN_2109489	-1.47	<0.0001	210164 at	N/A
LOC652493	652493	ILMN_1739508	-1.61	<0.0001	N/A	N/A
HBM	3042	ILMN_2091454	-1.62	<0.0001	240336 at	N/A
CD14	929	ILMN_2396444	-1.74	<0.0001	201743 at	N/A
ALAS2	212	ILMN_2367126	-1.76	<0.0001	211560 s at	N/A
HBB	3043	ILMN_2100437	-1.78	<0.0001	209116 x at	N/A
LOC642113	642113	ILMN_1652199	-1.79	<0.0001	N/A	N/A
AHSP	51327	ILMN_1696512	-1.84	<0.0001	219672 at	N/A
FCN1	2219	ILMN_1668063	-1.85	<0.0001	205237 at	N/A
CD48	962	ILMN_2061043	-1.85	<0.0001	204118 at	N/A
HBA2	3040	ILMN_2127842	-2.06	<0.0001	N/A	N/A
HBA1	3039	ILMN_3240144	-2.07	<0.0001	N/A	N/A

*Illumina microarray probe IDs.

†Fold change LSC⁺ compared to LSC⁻ GE profiles.‡Student's *t*-test *P* values for fold changes.

§Affymetrix microarray probeset IDs.

Extended Data Table 2 | Clinical characteristics of the TCGA AML cohort

Characteristic	TCGA AML cohort (n=183)	High LSC17 score subset (n=92)	Low LSC17 score subset (n=91)	P-value
Female Sex [n (%)]	85 (46.4)	39 (42.4)	46 (50.5)	0.33‡
Age at AML Diagnosis [years]				
median (range)	57 (18-88)	61 (21-88)	55 (18-82)	0.002†
PB WBC count at diagnosis (x10⁹/L)				
median (range)	16.8 (0.5-298.4)	12.1 (0.5-297.4)	26.1 (0.6-298.4)	0.06*
Blast % at diagnosis (x10⁹/L)				
Median BM Blasts (range)	72 (30-100)	73 (30-99)	72 (30-100)	0.97*
Median PB Blasts (range)	33 (0-98)	35.5 (0-98)	33 (0-97)	0.57*
AML subtypes [n (%)]				
APL	17 (9.29)	4 (4.35)	13 (14.3)	0.02
Non-APL	166 (90.7)	88 (95.7)	78 (85.7)	
Cytogenetic risk class at diagnosis [n (%)]¶				
Favorable	38 (20.8)	7 (7.61)	31 (34.1)	<0.001
Intermediate	105 (57.4)	53 (57.6)	52 (57.1)	
Adverse	40 (21.9)	32 (34.8)	8 (8.79)	
Karyotype [n (%)]	n=179	n=90	n=89	
Normal karyotype	83 (46.4)	36 (40)	47 (52.8)	0.11‡
Abnormal karyotype	96 (53.6)	54 (60)	42 (47.2)	
Molecular Risk at diagnosis [n (%)]	n=180	n=91	n=89	
Favorable	35 (19.4)	5 (5.49)	30 (33.7)	<0.001
Intermediate	98 (54.4)	50 (54.9)	48 (53.9)	
Adverse	47 (26.1)	36 (39.6)	11 (12.4)	
CN AML- <i>NPM1</i> [n (%)]	n=83	n=36	n=47	
<i>NPM1</i> mutation	42 (50.6)	18 (50)	24 (51.1)	1.00‡
No <i>NPM1</i> mutation	41 (49.4)	18 (50)	23 (48.9)	
CN AML- <i>FLT3</i>-ITD [n (%)]	n=83	n=36	n=47	
<i>FLT3</i> -ITD positive	23 (27.7)	16 (44.4)	7 (14.9)	0.005
<i>FLT3</i> -ITD negative	60 (72.3)	20 (55.6)	40 (85.1)	
Survival Parameters [days]				
Median Overall Survival	492	303	1029	<0.001§

*P value calculated using the Wilcoxon rank-sum test.

†P value calculated using the Student's t-test.

‡P value calculated using the Pearson's chi-squared test.

§P value calculated using the log-rank test.

||P value calculated using the Fisher's exact test.

¶Cytogenetic risk groups were defined by TCGA research network¹⁸.

Extended Data Table 3 | Clinical characteristics of the GSE12417 CN-AML cohorts

a

Characteristic	GSE12417 CN-AML cohort 1 (n=156)	High LSC17 score subset (n=78)	Low LSC17 score subset (n=78)	P-value
Female Sex [n (%)]	84 (53.8)	39 (50)	45 (57.7)	0.42‡
Age at AML Diagnosis [years]				
median (range)	57 (17-83)	61 (20-81)	54.5 (17-83)	0.03†
De novo vs. Secondary AML [n (%)]				
De novo	149 (95.5)	74 (94.9)	75 (96.2)	1.00
Secondary / t-AML	7 (4.49)	4 (5.13)	3 (3.85)	
PB WBC count at diagnosis (x10⁹/L)				
median (range)	36.2 (0.095-486)	45.3 (0.9-486)	30.6 (0.095-289)	0.18*
BM blast % at diagnosis (x10⁹/L)	n=153	n=77	n=76	
median (range)	85 (20-100)	90 (20-100)	80 (20-100)	0.04*
NPM1 [n (%)]				
NPM1 mutation	83 (53.2)	42 (53.8)	41 (52.6)	1.00‡
No NPM1 mutation	73 (46.8)	36 (46.2)	37 (47.4)	
FLT3-ITD [n (%)]				
FLT3-ITD positive	75 (48.1)	53 (67.9)	22 (28.2)	<0.001‡
FLT3-ITD negative	81 (51.9)	25 (32.1)	56 (71.8)	
Treatment Response [n (%)]				
Complete Response	94 (60.3)	37 (47.4)	57 (73.1)	0.001‡
No Response	62 (39.7)	41 (52.6)	21 (26.9)	
Survival Parameters [days]				
Median Overall Survival	294	223	not reached	<0.001§
Median Event-Free Survival	192 (n=153)	83 (n=76)	371 (n=77)	<0.001§
Median Relapse-Free Survival	384 (n=91)	178 (n=36)	627 (n=55)	0.001§

b

Characteristic	GSE12417 CN-AML cohort 2 (n=70)	High LSC17 score subset (n=35)	Low LSC17 score subset (n=35)	P-value
Female Sex [n (%)]	29 (41.4)	18 (51.4)	11 (31.4)	0.14
Age at AML Diagnosis [years]				
median (range)	62 (18-85)	62 (22-81)	62 (18-85)	0.15†
De novo vs. Secondary AML [n (%)]				
De novo	62 (88.6)	28 (80)	34 (97.1)	0.05
Secondary / t-AML	8 (11.4)	7 (20)	1 (2.86)	
PB WBC count at diagnosis (x10⁹/L)	n=68	n=34	n=34	
median (range)	15 (1-440.3)	14.0 (1-440.3)	17.8 (1-280)	0.85*
BM blast % at diagnosis (x10⁹/L)	n=67	n=33	n=34	
median (range)	80 (18-97)	80 (18-95)	87.5 (20-97)	0.26*
NPM1 [n (%)]				
NPM1 mutation	36 (51.4)	13 (37.1)	23 (65.7)	0.03
No NPM1 mutation	34 (48.6)	22 (62.9)	12 (34.3)	
FLT3-ITD [n (%)]				
FLT3-ITD positive	19 (27.1)	14 (40)	5 (14.3)	0.03
FLT3-ITD negative	51 (72.9)	21 (60)	30 (85.7)	
Treatment Response [n (%)]	n=68	n=33		
Complete Response	43 (63.2)	15 (45.5)	28 (80)	0.005
No Response	25 (36.8)	18 (54.5)	7 (20)	
Survival Parameters [days]				
Median Overall Survival	500	301	not reached	0.001§
Median Event-Free Survival	243 (n=69)	120 (n=34)	398	<0.001§
Median Relapse-Free Survival	368 (n=40)	183 (n=14)	not reached (n=26)	0.02§

*P value calculated using the Wilcoxon rank-sum test.

†P value calculated using the Student's t-test.

‡P value calculated using the Pearson's chi-squared test.

§P value calculated using the log-rank test.

||P value calculated using the Fisher's exact test.

Extended Data Table 4 | Multivariate survival analysis of LSC17 and LSC3 scores

a

Overall Survival	TCGA AML (n=183)*		TCGA AML RNA-Seq (n=166)*	
Covariate	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	2.50 (1.37-4.58)	0.002	1.91 (1.23-2.98)	0.003
Age	Stratifier§	N/A	1.03 (1.02-1.05)	<0.001
WBC count	1.01 (1.00-1.015)	0.004	1.01 (1.00-1.01)	<0.001
Favorable Cytogenetics	0.73 (0.36-1.49)	0.39	0.71 (0.37-1.36)	0.30
Adverse Cytogenetics	1.52 (0.80-2.89)	0.19	1.57 (1.02-2.41)	0.04
Covariate	Hazard Ratio (95% CI)†	P-value‡	TCGA CN-AML RNA-Seq (n=76)* Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	5.32 (1.27-22.3)	0.02	2.44 (1.25-4.77)	0.008
Age	Stratifier§	N/A	1.02 (1.00-1.046)	0.02
WBC count	1.02 (1.00-1.03)	0.01	1.01 (1.00-1.01)	0.002
<i>NPM1</i> Mutation	1.01 (0.26-3.90)	0.98	0.96 (0.53-1.74)	0.90
<i>FLT3</i> -ITD Mutation	5.23 (1.12-24.4)	0.03	1.28 (0.64-2.56)	0.48
Covariate	GSE12417 CN-AML Cohort 1 (n=156)* Hazard Ratio (95% CI)†	P-value‡	GSE12417 CN-AML Cohort 2 (n=68)* Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	2.45 (1.54-3.89)	<0.001	2.29 (1.08-4.84)	0.02
Age	1.02 (1.00-1.04)	0.009	1.03 (1.00-1.06)	0.04
WBC count	1.00 (1.00-1.00)	0.95	1.00 (1.00-1.00)	0.002
<i>NPM1</i> Mutation	0.72 (0.47-1.10)	0.13	0.52 (0.25-1.08)	0.08
<i>FLT3</i> -ITD Mutation	1.88 (1.17-3.02)	0.009	1.08 (0.47-2.48)	0.85
Secondary / t-AML	1.49 (0.65-3.42)	0.34	0.56 (0.17-1.83)	0.34
Covariate	PM AML (n=284)* Hazard Ratio (95% CI)†	P-value‡	PM CN-AML (n=85)* Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	2.49 (1.78-3.48)	<0.001	2.02 (1.04-3.92)	0.03
Age	1.00 (0.99-1.01)	0.25	1.01 (0.98-1.03)	0.40
WBC count	1.00 (1.00-1.00)	0.003	1.00 (0.99-1.00)	0.11
<i>NPM1</i> Mutation	N/A	N/A	0.44 (0.22-0.92)	0.02
<i>FLT3</i> -ITD Mutation	N/A	N/A	1.96 (0.97-3.95)	0.05
Favorable Cytogenetics	0.46 (0.26-0.79)	0.005	N/A	N/A
Adverse Cytogenetics	1.96 (1.33-2.91)	<0.001	N/A	N/A
Secondary / t-AML	2.39 (1.61-3.54)	<0.001	2.63 (1.15-6.01)	0.02

b

Overall Survival	GSE15434 CN-LMR (n=70)*		PM CN-LMR (n=29)*	
Covariate	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC3 Score	8.49 (2.46-29.3)	<0.001	6.30 (1.22-32.3)	0.02
Age	0.99 (0.94-1.04)	0.81	0.99 (0.93-1.05)	0.84
WBC count	N/A	N/A	1.01 (1.00-1.02)	0.02
Secondary / t-AML	1.19 (0.34-4.22)	0.78	11.5 (1.42-93.5)	0.02

t-AML, therapy-associated AML.

*Number of patients with full clinical annotation.

†95% confidence interval.

‡P value calculated using the Wald test.

§Age violated the proportional hazards assumption.

Extended Data Table 5 | The LSC17 score refines genomic classifications

a

TCGA AML (n=183)*	Univariate Analysis		Multivariate Analysis 1		Multivariate Analysis 2 (P<0.001)	
	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	N/A	N/A	Not included in model	N/A	3.08 (1.56-6.06)	0.001
PML-RARA Mutation	0.30 (0.12-0.73), n=15	0.008	0.43 (0.10-1.74)	0.23	0.34 (0.08-1.47)	0.15
MYH11-CBFB Mutation	0.33 (0.12-0.90), n=10	0.03	0.40 (0.08-1.96)	0.26	0.27 (0.05-1.38)	0.11
FLT3 in-frame Mutation	5.98 (2.12-16.8), n=4	<0.001	4.78 (0.73-31.3)	0.10	7.75 (0.90-66.2)	0.06
DNMT3A Mutation	1.58 (1.07-2.32), n=45	0.02	1.74 (0.95-3.17)	0.07	1.92 (1.02-3.58)	0.04
RUNX1 Mutation	1.79 (1.05-3.03), n=17	0.03	1.44 (0.66-3.17)	0.35	1.08 (0.48-2.45)	0.84
TP53 Mutation	3.61 (2.09-6.22), n=16	<0.001	2.30 (0.92-5.71)	0.07	1.71 (0.66-4.38)	0.26
Age	N/A	N/A	Stratifier§	N/A	Stratifier§	N/A
WBC count	N/A	N/A	1.00 (1.00-1.01)	0.006	1.01 (1.00-1.01)	0.005
Favorable Cytogenetics	N/A	N/A	1.32 (0.43-4.06)	0.62	2.33 (0.68-7.92)	0.17
Adverse Cytogenetics	N/A	N/A	1.61 (0.78-3.31)	0.19	1.40 (0.65-3.02)	0.39

b

Overall Survival TCGA AML (n=183)*	Multivariate Analysis 1		Multivariate Analysis 2 (P=0.001)	
	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 score	Not included in model	N/A	2.85 (1.47-5.52)	0.002
Age	Stratifier§	N/A	Stratifier§	N/A
WBC count	1.00 (1.00-1.01)	0.02	1.00 (1.00-1.01)	0.01
CEBPA ^{haeic}	0.86 (0.11-6.57)	0.89	1.45 (0.18-11.5)	0.72
Chromatin-spliceosome	1.80 (0.54-5.98)	0.33	1.00 (0.28-3.54)	0.99
IDH2 ¹⁷²	<0.001 (0.00-inf)	0.99	<0.001 (0.00-inf)	0.99
inv(16)	1.08 (0.23-5.03)	0.91	0.82 (0.17-3.85)	0.80
MLL fusion	2.19 (0.51-9.35)	0.28	1.77 (0.39-7.89)	0.45
No class-defining drivers	2.19 (0.60-7.98)	0.23	1.54 (0.40-5.87)	0.52
No driver mutations	1.20 (0.16-8.60)	0.85	0.71 (0.09-5.34)	0.74
NPM1 mutation	2.21 (0.70-6.93)	0.17	1.65 (0.51-5.30)	0.39
t(8;21)	1.52 (0.21-10.8)	0.67	2.22 (0.29-16.5)	0.43
TP53-aneuploidy	4.53 (1.34-15.3)	0.01	2.12 (0.58-7.73)	0.25
2+ classes	3.26 (0.48-21.9)	0.22	1.70 (0.24-11.8)	0.59

c

Overall Survival TCGA AML (n=33)*	Multivariate Analysis 1		Multivariate Analysis 2 (P=0.01)	
	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 score	Not included in model	N/A	3.33 (1.30-8.51)	0.01
Age	1.05 (1.02-1.09)	<0.001	1.04 (1.01-1.08)	0.006
WBC count	1.01 (1.00-1.02)	0.02	1.01 (1.00-1.02)	0.006
No driver mutations	0.54 (0.12-2.41)	0.42	0.34 (0.07-1.53)	0.16
2+ classes	1.84 (0.51-6.61)	0.34	1.19 (0.32-4.37)	0.78

b, Genomic classes were compared to t(15;17) in CPH models. **c**, Genomic classes were compared to the no class-defining drivers category²¹ in CPH models. Patient scores above or below the median score of the entire TCGA AML cohort were designated as high or low LSC17 score, respectively.

*Number of patients with full clinical annotation.

†95% confidence interval.

‡P value was calculated using the Wald test.

§Age violated the proportional hazards assumption.

||P value was calculated using the likelihood ratio test to assess model improvement by including LSC17.

Extended Data Table 6 | The LSC17 score improves survival association compared to other LSC signatures

Overall Survival	GSE12417 CN-AML Cohort 1		GSE12417 CN-AML Cohort 2		TCGA AML	
	Univariate Analysis (n=156)*		Univariate Analysis (n=70)*		Univariate Analysis (n=183)*	
Covariate	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High SDPC Score	2.26 (1.50-3.40)	<0.001	1.43 (0.78-2.63)	0.24	0.96 (0.68-1.35)	0.82
High IFPC Score	2.18 (1.45-3.27)	<0.001	1.65 (0.89-3.06)	0.10	1.19 (0.85-1.68)	0.30
High Jung et al. Score	2.35 (1.56-3.54)	<0.001	1.45 (0.78-2.67)	0.23	2.02 (1.42-2.87)	<0.001
High Gentles et al. Score	2.15 (1.43-3.22)	<0.001	1.21 (0.66-2.23)	0.53	1.68 (1.18-2.38)	0.003
Covariate	Multivariate Analysis (P<0.001)§		Multivariate Analysis (P=0.02)§		Multivariate Analysis (P=0.001)§	
	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	2.31 (1.42-3.76)	<0.001	2.33 (1.10-4.91)	0.02	2.67 (1.45-4.92)	0.001
High SDPC Score	1.20 (0.73-1.98)	0.46	0.53 (0.23-1.23)	0.14	0.71 (0.40-1.26)	0.24
Age	1.02 (1.00-1.03)	0.01	1.03 (1.00-1.07)	0.02	Stratifier§	N/A
WBC count	1.00 (1.00-1.00)	0.92	1.00 (1.00-1.00)	<0.001	1.01 (1.00-1.01)	0.006
Favorable Cytogenetics	N/A	N/A	N/A	N/A	0.84 (0.39-1.77)	0.64
Adverse Cytogenetics	N/A	N/A	N/A	N/A	1.70 (0.86-3.36)	0.12
Secondary / t-AML	1.54 (0.67-3.55)	0.30	0.50 (0.15-1.63)	0.25	N/A	N/A
FLT3-ITD Mutation	1.81 (1.11-2.94)	0.01	1.21 (0.51-2.83)	0.65	N/A	N/A
NPM1 Mutation	0.77 (0.48-1.23)	0.27	0.35 (0.14-0.87)	0.02	N/A	N/A
Covariate	Multivariate Analysis (P<0.001)§		Multivariate Analysis (P=0.03)§		Multivariate Analysis (P=0.003)§	
	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	2.31 (1.44-3.71)	<0.001	2.17 (1.02-4.61)	0.04	2.47 (1.30-4.69)	0.005
High IFPC Score	1.28 (0.81-2.04)	0.28	1.40 (0.70-2.81)	0.33	1.03 (0.58-1.83)	0.91
Age	1.02 (1.00-1.03)	0.01	1.03 (1.00-1.06)	0.03	Stratifier§	N/A
WBC count	1.00 (1.00-1.00)	0.99	1.00 (1.00-1.00)	0.002	1.01 (1.00-1.01)	0.004
Favorable Cytogenetics	N/A	N/A	N/A	N/A	0.73 (0.35-1.49)	0.38
Adverse Cytogenetics	N/A	N/A	N/A	N/A	1.51 (0.79-2.89)	0.20
Secondary / t-AML	1.52 (0.66-3.48)	0.32	0.57 (0.17-1.86)	0.35	N/A	N/A
FLT3-ITD Mutation	1.75 (1.07-2.85)	0.02	0.95 (0.40-2.25)	0.92	N/A	N/A
NPM1 Mutation	0.71 (0.46-1.09)	0.12	0.52 (0.25-1.08)	0.08	N/A	N/A
Covariate	Multivariate Analysis (P=0.001)§		Multivariate Analysis (P=0.01)§		Multivariate Analysis (P=0.01)§	
	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	2.30 (1.38-3.85)	0.001	3.24 (1.29-8.13)	0.01	2.25 (1.17-4.29)	0.01
High Jung et al. Score	1.14 (0.69-1.88)	0.59	0.59 (0.26-1.33)	0.20	1.37 (0.70-2.67)	0.35
Age	1.02 (1.00-1.04)	0.01	1.03 (1.00-1.06)	0.04	Stratifier§	N/A
WBC count	1.00 (1.00-1.00)	0.84	1.00 (1.00-1.00)	0.001	1.01 (1.00-1.01)	0.003
Favorable Cytogenetics	N/A	N/A	N/A	N/A	0.71 (0.34-1.46)	0.35
Adverse Cytogenetics	N/A	N/A	N/A	N/A	1.41 (0.73-2.74)	0.30
Secondary / t-AML	1.48 (0.64-3.40)	0.35	0.49 (0.15-1.63)	0.25	N/A	N/A
FLT3-ITD Mutation	1.82 (1.11-2.97)	0.01	1.01 (0.43-2.36)	0.96	N/A	N/A
NPM1 Mutation	0.73 (0.47-1.12)	0.15	0.50 (0.24-1.03)	0.06	N/A	N/A
Covariate	Multivariate Analysis (P=0.001)§		Multivariate Analysis (P=0.01)§		Multivariate Analysis (P=0.009)§	
	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡	Hazard Ratio (95% CI)†	P-value‡
High LSC17 Score	2.17 (1.31-3.58)	0.002	2.56 (1.14-5.71)	0.02	2.27 (1.21-4.28)	0.01
High Gentles et al. Score	1.33 (0.83-2.11)	0.23	0.74 (0.34-1.59)	0.44	1.37 (0.79-2.38)	0.25
Age	1.02 (1.00-1.04)	0.008	1.03 (0.99-1.06)	0.05	Stratifier§	N/A
WBC count	1.00 (1.00-1.00)	0.74	1.00 (1.00-1.00)	0.002	1.00 (1.00-1.01)	0.003
Favorable Cytogenetics	N/A	N/A	N/A	N/A	0.70 (0.34-1.43)	0.33
Adverse Cytogenetics	N/A	N/A	N/A	N/A	1.41 (0.73-2.71)	0.29
Secondary / t-AML	1.49 (0.65-3.41)	0.34	0.48 (0.14-1.68)	0.25	N/A	N/A
FLT3-ITD Mutation	1.80 (1.12-2.91)	0.01	1.17 (0.49-2.79)	0.71	N/A	N/A
NPM1 Mutation	0.74 (0.48-1.13)	0.16	0.48 (0.22-1.03)	0.06	N/A	N/A

SDPC, surface-defined primitive cells²²; IFPC, inferred functionally primitive cells²².

*Number of patients with full clinical annotation.

†95% confidence interval.

‡P value was calculated using the Wald test.

§Age violated the proportional hazards assumption.

||P value was calculated using the likelihood ratio test to assess model improvement by including LSC17.

Extended Data Table 7 | Clinical characteristics of the PM AML and GSE15434 CN-LMR AML cohorts

a

Characteristic	PM AML (n=307)	High LSC17 score (n=154)	Low LSC17 score (n=153)	P-value
Female Sex [n (%)]	148 (48.2)	77 (50)	71 (46.4)	0.60†
Age at AML Diagnosis [years]				
median (range)	52 (18-81)	56 (18-81)	49 (20-81)	<0.001†
De novo vs. Secondary AML [n (%)]				
De novo	268 (87.3)	130 (84.4)	138 (90.2)	0.17‡
Secondary / t-AML	39 (12.7)	24 (15.6)	15 (9.8)	
PB WBC count at diagnosis (x10⁹/L)				
median (range)	17.6 (0.7-399)	12.2 (0.7-212)	26.8 (1.6-399)	<0.001*
BM blast % at diagnosis (x10⁹/L)	n=284	n=142	n=142	
median (range)	80 (10-98)	80 (10-98)	80 (16-95)	0.05*
AML subtypes [n (%)]	n=251	n=126	n=125	
APL	12 (4.78)	7 (5.56)	5 (4)	0.77
Non-APL	239 (95.2)	119 (94.4)	120 (96)	
Karyotype [n (%)]	n=284	n=141	n=143	
Normal karyotype	141 (49.6)	61 (43.3)	80 (55.9)	0.04‡
Abnormal karyotype	143 (50.4)	80 (56.7)	63 (44.1)	
MRC Cytogenetic risk class at diagnosis [n (%)]	n=284	n=141	n=143	
Favorable	48 (16.9)	13 (9.22)	35 (24.5)	<0.001
Intermediate	196 (69)	91 (64.5)	105 (73.4)	
Adverse	40 (14.1)	37 (26.2)	3 (2.1)	
CN AML- NPM1 [n (%)]	n=87	n=29	n=58	
NPM1 mutation	48 (55.2)	9 (31)	39 (67.2)	0.002
No NPM1 mutation	39 (44.8)	20 (69)	19 (32.8)	
CN AML- FLT3-ITD [n (%)]	n=95	n=34	n=61	
FLT3-ITD positive	23 (24.2)	8 (23.5)	15 (24.6)	1.00
FLT3-ITD negative	72 (75.8)	26 (76.5)	46 (75.4)	
Treatment Response [n (%)]	n=306	n=153	n=153	
Complete Remission	223 (72.9)	85 (55.6)	138 (90.2)	<0.001
No Response	83 (27.1)	68 (44.4)	15 (9.8)	
Survival Parameters [n (%)]				
Median Overall Survival	671	400	2035	<0.001§
Median Event-Free Survival	301	161	541	<0.001§
Median Relapse-Free Survival	378 (n=267)	265 (n=118)	689 (n=149)	<0.001§

b

Characteristic	GSE15434 CN-LMR cohort (n=70)	High LSC3 score subset (n=35)	Low LSC3 score subset (n=35)	P-value
Female Sex [n (%)]	34 (48.6)	16 (45.7)	18 (51.4)	0.81‡
Age at AML Diagnosis [years]				
median (range)	54 (30-83)	57 (36-83)	51 (30-75)	0.08†
De novo vs. Secondary AML [n (%)]				
De novo	65 (92.9)	32 (91.4)	33 (94.3)	1.00
Secondary / t-AML	5 (7.14)	3 (8.57)	2 (5.71)	
PB WBC count at diagnosis (x10⁹/L)	n=45	n=21	n=24	
median (range)	17.9 (0.9-365)	29.4 (1.6-365)	14.4 (0.9-86)	0.11*
Blast % at diagnosis (x10⁹/L)	n=69,67	n=35,33	n=34	
Median BM Blasts (range)	69 (0-95)	75 (0-95)	65.5 (14-95)	0.11*
Median PB Blasts (range)	18 (0-94)	60 (0-94)	9 (0-90)	0.003*
Karyotype [n (%)]				
Normal karyotype	70 (100)	35 (100)	35 (100)	1.00
Abnormal karyotype	0 (0)	0 (0)	0 (0)	
Treatment Response [n (%)]	n=39	n=20	n=19	
Complete Response	35 (89.7)	18 (90)	17 (89.5)	1.00
No Response	4 (10.3)	2 (10)	2 (10.5)	
CN AML- NPM1 [n (%)]				
NPM1 mutation	70 (100)	35 (100)	35 (100)	1.00
No NPM1 mutation	0 (0)	0 (0)	0 (0)	
CN AML- FLT3-ITD [n (%)]				
FLT3-ITD positive	0 (0)	0 (0)	0 (0)	1.00
FLT3-ITD negative	70 (100)	35 (100)	35 (100)	
Survival Parameters [days]				
Median Overall Survival	1767	679	not reached	<0.001§

*P value calculated using the Wilcoxon rank-sum test.

†P value calculated using the Student's t-test.

‡P value calculated using the Pearson's chi-squared test.

§P value calculated using the log-rank test.

||P value calculated using the Fisher's exact test.

Extended Data Table 8 | Clinical characteristics and multivariate survival analysis of the ALFA-0701 AML cohort

a

Characteristic	ALFA-0701 Trial Cohort (n=192)	High LSC17 score subset (n=96)	Low LSC17 score subset (n=96)	P-value
Female Sex [n (%)]	98 (51)	45 (0.47)	53 (0.55)	0.31‡
Age at AML Diagnosis [years]				
median (range)	62.1 (50.1-70.8)	61.8 (50.1-70.8)	62.3 (50.5-70.7)	0.89†
De novo vs. Secondary AML [n (%)]				
De novo	192 (100)	96 (100)	96 (100)	N/A
Secondary / t-AML	0 (0)	0 (0)	0 (0)	
PB WBC count at diagnosis (x10⁹/L)	n=191	n=95		
median (range)	5.1 (0.15-210.6)	3.82 (0.5-210.6)	8.8 (0.15-187)	0.01*
Treatment arm (GO or standard)				
Gemtuzumab Ozogamicin	98 (0.51)	47 (0.49)	51 (0.53)	0.66‡
Standard treatment	94 (0.49)	49 (0.51)	45 (0.47)	
Cytogenetic risk class at diagnosis [n (%)]¶	n=175	n=91	n=84	
Favorable	4 (0.02)	0 (0)	4 (0.04)	<0.001§
Intermediate	129 (0.73)	59 (0.64)	70 (0.83)	
Adverse	42 (0.24)	32 (0.35)	10 (0.12)	
Karyotype [n (%)]	n=173	n=91	n=82	
Normal karyotype	99 (0.57)	44 (0.48)	55 (0.67)	0.02‡
Abnormal karyotype	74 (0.42)	47 (0.51)	27 (0.33)	
AML subtypes [n (%)]				
APL	0 (0)	0 (0)	0 (0)	N/A
Non-APL	192 (100)	96 (100)	96 (100)	
CN AML- NPM1 [n (%)]	n=99	n=44	n=55	
NPM1 mutation	44 (0.44)	16 (0.36)	28 (0.51)	0.001‡
No NPM1 mutation	55 (0.55)	28 (0.63)	27 (0.49)	
CN AML- FLT3-ITD [n (%)]	n=99	n=44	n=55	
FLT3-ITD positive	24 (0.24)	17 (0.38)	7 (0.12)	0.08
FLT3-ITD negative	75 (0.75)	27 (0.61)	48 (0.87)	
Treatment Response [n (%)]				
Complete Response	132 (0.68)	61 (0.63)	71 (0.74)	0.16‡
No Response	60 (0.31)	35 (0.36)	25 (0.26)	
Events [n (%)]	n=145	n=65	n=80	
Relapse	90 (0.62)	46 (0.70)	44 (0.55)	0.07‡
Survival Parameters [days]				
Median Overall Survival	666	462	1387	<0.001§
Median Event-Free Survival	330	241	532	<0.001§
Median Relapse-Free Survival	504 (n=145)	362 (n=65)	902 (n=80)	<0.001§

b

ALFA-0701 Low LSC17 Score Subset	Event-Free Survival (n=90)#		Relapse-Free Survival (n=80)#	
Covariate	Hazard Ratio (95% CI) ☆	P-value**	Hazard Ratio (95% CI) ☆	P-value**
GO vs. Standard Treatment	0.34 (0.19-0.63)	<0.001	0.42 (0.21-0.83)	0.01
Age	1.04 (0.99-1.10)	0.09	1.03 (0.97-1.09)	0.31
WBC count	1.00 (0.99-1.01)	0.11	1.00 (0.99-1.01)	0.10
Favorable Cytogenetics	1.29 (0.39-4.23)	0.66	1.26 (0.30-5.31)	0.75
Adverse Cytogenetics	1.98 (0.82-4.77)	0.12	0.44 (0.06-3.31)	0.43

*P value calculated using the Wilcoxon rank-sum test.

†P value calculated using the Student's t-test.

‡P value calculated using the Pearson's chi-squared test.

§P value calculated using the log-rank test.

||P value calculated using the Fisher's exact test.

¶Cytogenetic risk groups were defined by ALFA-0701 investigators²⁸.

#Number of patients with full clinical annotation.

☆95% confidence interval.

**P value was calculated using the Wald test.

Zika virus infection damages the testes in mice

Jennifer Govero^{1*}, Prabakaran Esakky^{2*}, Suzanne M. Scheaffer², Estefania Fernandez³, Andrea Drury², Derek J. Platt⁴, Matthew J. Gorman³, Justin M. Richner¹, Elizabeth A. Caine¹, Vanessa Salazar¹, Kelle H. Moley^{2,5} & Michael S. Diamond^{1,3,4,6}

Infection of pregnant women with Zika virus (ZIKV) can cause congenital malformations including microcephaly, which has focused global attention on this emerging pathogen¹. In addition to transmission by mosquitoes, ZIKV can be detected in the seminal fluid of affected males for extended periods of time and transmitted sexually². Here, using a mouse-adapted African ZIKV strain (Dakar 41519), we evaluated the consequences of infection in the male reproductive tract of mice. We observed persistence of ZIKV, but not the closely related dengue virus (DENV), in the testis and epididymis of male mice, and this was associated with tissue injury that caused diminished testosterone and inhibin B levels and oligospermia. ZIKV preferentially infected spermatogonia, primary spermatocytes and Sertoli cells in the testis, resulting in cell death and destruction of the seminiferous tubules. Less damage was caused by a contemporary Asian ZIKV strain (H/PF/2013), in part because this virus replicates less efficiently in mice. The extent to which these observations in mice translate to humans remains unclear, but longitudinal studies of sperm function and viability in ZIKV-infected humans seem warranted.

We and others have observed that infection of male adult mice with ZIKV results in infection of the testes^{3,4}, which is consistent with observed male-to-female^{5,6} and male-to-male⁷ sexual transmission in humans. To address the effects of infection on the male reproductive tract, we performed a longitudinal study in wild-type C57BL/6 mice infected with ZIKV (strains H/PF/2013 (French Polynesia 2013) or mouse-adapted Dakar 41519 (Senegal 1984)) or DENV (serotype 2, strain D2S20). Because ZIKV and DENV do not efficiently antagonize type I interferon (IFN) signalling in mice compared to humans⁸, animals were treated with a single dose of IFN α and IFN β receptor 1 (Ifnar1)-blocking monoclonal antibody to facilitate infection and dissemination. When wild-type mice were treated with an isotype-control antibody instead and then infected, ZIKV RNA did not accumulate in the testes (Fig. 1a).

In the presence of the anti-Ifnar1 antibody, high levels of viral RNA (10^5 – 10^8 focus-forming unit (FFU) equivalents per g or ml) and infectious virus (up to 10^8 plaque-forming units (PFU) per g or ml) were detected in the testis, epididymis and the fluid collected from the epididymis within seven days of infection with either of the two ZIKV strains but not DENV (Fig. 1a–c). ZIKV-Dakar replicated to higher levels than ZIKV-H/PF/2013, which is consistent with the enhanced virulence of ZIKV-Dakar in wild-type mice³. Notably, ZIKV RNA and infectious virus were also detected in mature sperm collected from the epididymis (Fig. 1b, c, and Extended Data Fig. 1). At day 7 after inoculation, ZIKV-infected testes appeared similar in size to uninfected testes from age-matched mice and had equivalent weights (Fig. 1d, e). Histological analysis of ZIKV-infected testis and epididymis at day 7 revealed no apparent differences in architecture (Fig. 1f and Extended Data Fig. 2). However, staining for CD45 (a pan-leukocyte marker) was observed in testis sections only from ZIKV-infected animals, with

CD45⁺ cells localizing to the interstitium between the seminiferous tubules (Fig. 1g, column 1). The blood–testis barrier (BTB) remained intact at day 7 after infection, as shown by equivalent staining of the ETV5 transcription factor (which mediates BTB function and testicular immune privilege⁹) in Sertoli and germ cells in sections from uninfected and ZIKV-infected mice (Fig. 1g, column 2). Furthermore, there was no CD45 staining on the seminiferous tubular side of the BTB, near the TRA98⁺ germ cells or spermatogonia (Fig. 1g, column 1). A similar pattern of CD45 staining in the testicular interstitium and epididymal epithelium was described in patients infected with HIV¹⁰; indeed, we also observed scattered CD45⁺ cells in the epididymal epithelium of ZIKV-infected mice (Fig. 1g, column 5). However, at day 7, F4/80⁺ macrophages were not apparent in the testicular interstitium or the luminal epithelium of the epididymis of ZIKV-infected mice (Fig. 1g, columns 3 and 4).

To determine which cells were targeted by ZIKV, we performed *in situ* hybridization (ISH) for viral RNA at day 7 after infection. In the testis, ZIKV RNA was evident in spermatogonia, primary spermatocytes and the trophic, inhibin B-producing Sertoli cells (Fig. 1h, left), with relative sparing of the androgen-producing Leydig cells. In the cauda epididymis, mature sperm in the lumen stained strongly for ZIKV RNA (Fig. 1h, right) as did sperm cells collected from the epididymis (Extended Data Fig. 1).

We followed the consequences of ZIKV infection of the male reproductive tract over time. At day 14 after inoculation, high levels of ZIKV RNA persisted in the testis, epididymis, the fluid from the epididymis and mature sperm of most mice (Fig. 2a). In ZIKV-Dakar-infected animals, there was a noticeable decrease in testis size and weight (Fig. 2b, c). In comparison, no noticeable infection by DENV was observed in the testis at this time point (Extended Data Fig. 3a). Histological analysis of the ZIKV-infected testis at day 14 showed damage to the architecture of the seminiferous tubules with loss of the central ductal lumen (Fig. 2d). This was associated with decreased numbers of TRA98⁺ germ cells and Lina28a⁺ type A and B spermatogonia, morphological abnormalities of GATA4⁺ Sertoli cells and detachment of Sertoli cells from the basement membrane (Fig. 2e and Extended Data Fig. 2). In some regions, large numbers of CD45⁺ leukocytes were observed, suggesting substantial inflammatory cell infiltration (Fig. 2d, left; e, column 1). The absence of ETV5⁺ cells at this time point indicates loss of integrity of the BTB, which could explain the extent of interstitial inflammation and F4/80⁺ macrophages in the affected testis. The epididymis of ZIKV-infected animals also showed tissue injury at day 14, as indicated by constriction of the epididymal lumen, thickening of inter-luminal tissue and accumulation of sperm interspersed with necrotic bodies (Fig. 2d, e, right). ISH at day 14 showed progressive evidence of ZIKV RNA in cells of the testis, in the mature luminal sperm and on cilia layering the inner lumen of the epididymis, similar to day 7 (Fig. 2f).

High levels of viral RNA persisted in tissues of the male reproductive tract at 21 days after ZIKV-Dakar inoculation (Fig. 3a), and this

¹Department of Medicine, Washington University School of Medicine, Saint Louis, Missouri 63110, USA. ²Department of Obstetrics and Gynecology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA. ³Department of Pathology and Immunology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA. ⁴Department of Molecular Microbiology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA. ⁵Department of Cell Biology and Physiology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA. ⁶The Center for Human Immunology and Immunotherapy Programs, Washington University School of Medicine, Saint Louis, Missouri 63110, USA.

*These authors contributed equally to this work.

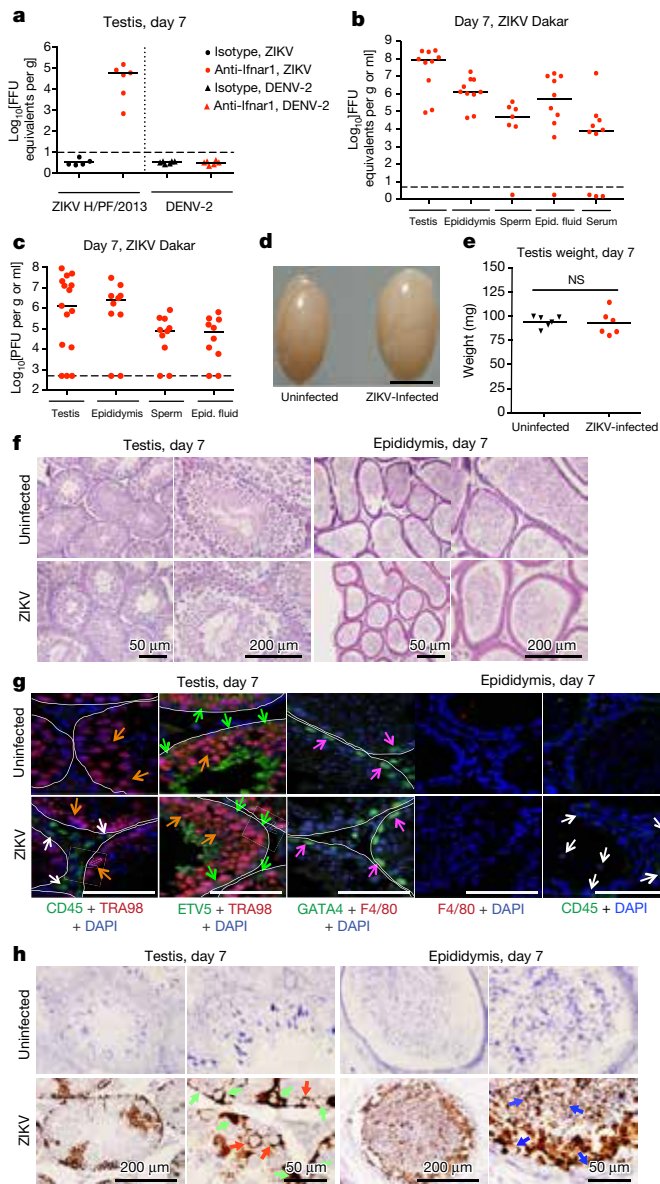


Figure 1 | ZIKV infection of the testis and epididymis at day 7. **a–c**, Seven-week-old wild-type mice were treated with an isotype control (**a**) or anti-Ifnar1 mouse antibody (2 mg (**a**) or 0.5 mg (**b**, **c**)) at day –1 before subcutaneous inoculation with 10^3 FFU of ZIKV-H/PF/2013 (**a**), 10^6 FFU of DENV-2 (**a**), or 10^6 FFU of mouse-adapted ZIKV-Dakar (**b**, **c**). Tissues and cells were collected at day 7 after infection and analysed for viral RNA by qRT-PCR (**a**, **b**) or for infectious virus by plaque assay (**c**). Dashed lines indicate limit of detection. Results are pooled from two or three independent experiments and each symbol represents data from an individual mouse. Bars indicate median values. Viral RNA was normalized to a standard curve from RNA isolated from infectious virus. **d**, A representative image of testes from uninfected and ZIKV-Dakar-infected mice at day 7; scale bar, 2 mm. **e**, Weight of testes from uninfected and ZIKV-infected mice at day 7. Results are pooled from two independent experiments. Mean values were not statistically different (NS; unpaired *t*-test). **f–h**, Histological, immunohistochemical and ISH analysis of testis (left) and epididymis (right) collected from uninfected or ZIKV-infected animals. **f**, Haematoxylin and eosin staining. **g**, Immunofluorescence staining of uninfected or ZIKV-infected testis and epididymis tissue sections with antibodies to CD45 (pan-leukocyte), TRA98 (germ cells), ETV5 (BTB), GATA4 (Sertoli cells) and F4/80 (macrophages). Arrows indicate staining for leukocytes (white), germ cells (orange), Sertoli cells (magenta) and BTB (green). White lines demarcate tubules of seminiferous epithelium. **h**, ISH with a ZIKV-specific probe. Arrows indicate cells positive for ZIKV RNA (spermatogonia and primary spermatocytes (red), Sertoli cells (green) and epididymis luminal sperm (blue)). The images in **f–h** are representative of several independent experiments. Scale bars, as indicated (**f**, **h**) and 50 μm (**g**).

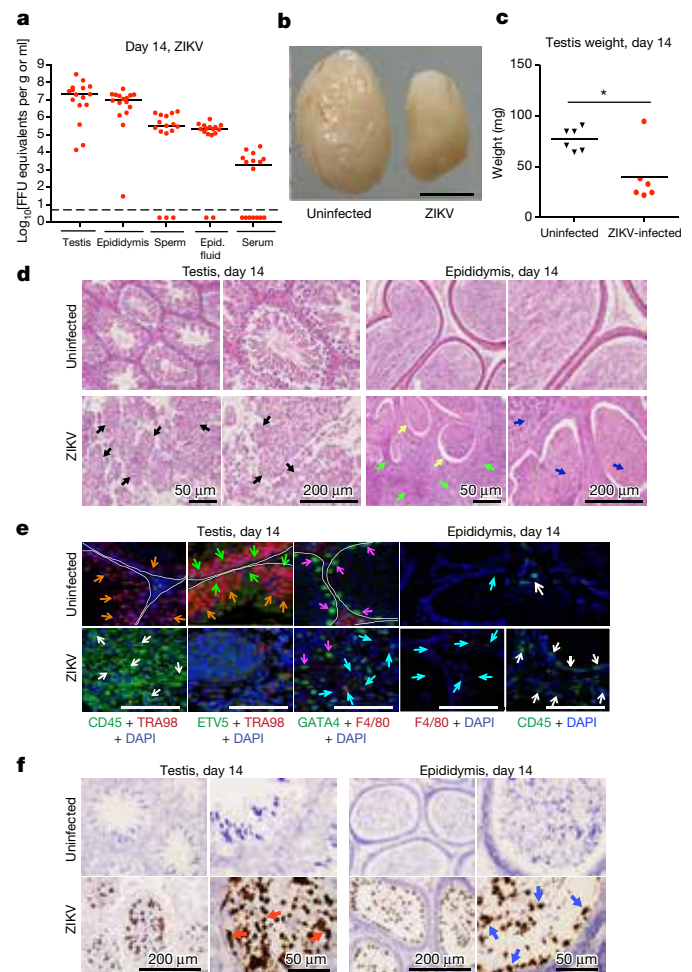


Figure 2 | ZIKV infection of the testis and epididymis at day 14. **a**, **b**, Seven-week-old wild-type mice were treated with 0.5 mg of anti-Ifnar1 at day –1 before subcutaneous inoculation of mouse-adapted ZIKV-Dakar. Tissues and cells were collected at day 14 and analysed for viral RNA by qRT-PCR (**a**). Dashed lines indicate limit of detection. Results are pooled from three independent experiments. Bars indicate median values. **b**, A representative image of testes from uninfected and ZIKV-infected mice at day 14; scale bar, 2 mm. **c**, Weight of testes from uninfected and ZIKV-infected mice at day 14. Results are pooled from two independent experiments (**P* < 0.05, Mann–Whitney test). **d–f**, Histological, immunohistochemical and ISH analysis of testis (left) and epididymis (right) collected from uninfected or ZIKV-infected animals. **d**, Haematoxylin and eosin staining. Arrows denote involution of seminiferous tubules in the testis (black), constricted epididymal lumens (yellow) with a mass of residual sperm (blue) and thickened epithelium (green). **e**, Immunofluorescence staining of uninfected or ZIKV-infected testis and epididymis tissues as described in Fig. 1. Arrows indicate staining for leukocytes (white), germ cells (orange), Sertoli cells (magenta), BTB (green) and macrophages (cyan). White lines demarcate tubules in the seminiferous epithelium. **f**, ISH. Arrows indicate cells positive for ZIKV RNA (testicular cells (red) and epididymis luminal sperm and cilia on the inner layer of epididymal epithelium (blue)). The images in **d–f** are representative of several independent experiments. Scale bars, as indicated (**d**, **f**) and 50 μm (**e**).

was associated with a loss of tissue architecture. Involution of the testis was observed, indicated by their noticeably reduced size and weight (Fig. 3b, c). Histological analysis revealed almost complete destruction of the seminiferous epithelium with constricted tubules after ZIKV infection (Fig. 3d). The populations of spermatogonia, Sertoli cells and 3β -HSD⁺ Leydig cells were markedly diminished, and this was associated with persistent CD45⁺ leukocyte infiltration (Fig. 3e and Extended Data Fig. 2). In the epididymis, ZIKV infection resulted in constriction of the lumen with a mass of residual sperm that was interspersed

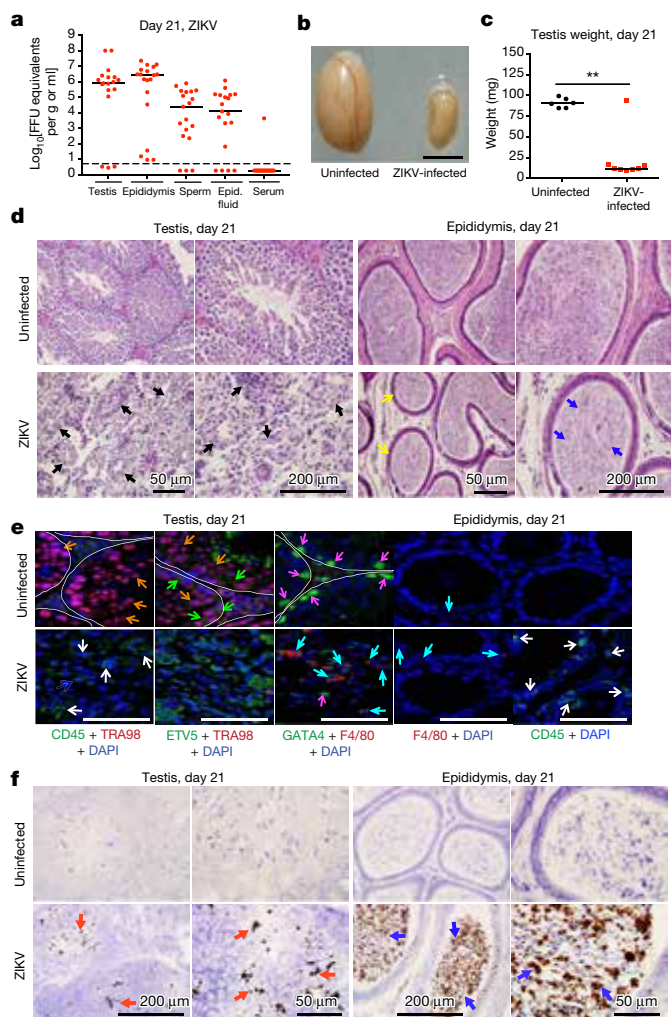


Figure 3 | ZIKV infection of the testis and epididymis at day 21.

a, Seven-week-old wild-type mice were treated with 0.5 mg of anti-Ifnar1 at day -1 before subcutaneous inoculation of mouse-adapted ZIKV-Dakar. Tissues and cells were collected at day 21 after infection and analysed for viral RNA by qRT-PCR. Dashed lines indicate limit of detection. Results are pooled from two independent experiments. Bars indicate median values. **b**, A representative image of testes from uninfected and ZIKV-infected mice at day 21; scale bar, 2 mm. **c**, Weight of testes from uninfected and ZIKV-infected mice at day 21. Results are pooled from two independent experiments (**P* < 0.05, Mann-Whitney test). **d**, Histological analysis of testis (left) and epididymis (right) collected from uninfected or ZIKV-infected animals stained with haematoxylin and eosin. Arrows indicate involution of seminiferous tubules in the testis (black), shrunken epididymal lumens (yellow) with a mass of residual sperm (blue). **e**, Immunofluorescence staining of uninfected or ZIKV-infected testis and epididymis tissues as described in Figs. 1, 2. Arrows indicate staining for leukocytes (white), germ cells (orange), Sertoli cells (magenta), BTB (green) and macrophages (cyan). White lines demarcate tubules in the seminiferous epithelium. **f**, ISH. Arrows indicate cells positive for ZIKV RNA (testicular cells (red) and epididymal luminal sperm (blue)). The images in **d**–**f** are representative of several independent experiments. Scale bars, as indicated (**d**, **f**) and 50 μ m (**e**).

with necrotic bodies (Fig. 3d). ISH showed viral RNA in remaining testicular cells of the damaged testis and in the luminal sperm of the infected epididymis (Fig. 3f). Damage to the seminiferous tubules in the testis, albeit at lower levels, was also observed in mice infected with the epidemic ZIKV-H/PF/2013 strain at day 28 after infection (Extended Data Fig. 3b, c).

The RNA ISH analysis suggested that Sertoli cells were targeted by ZIKV in the testis. Sertoli cells provide a trophic function

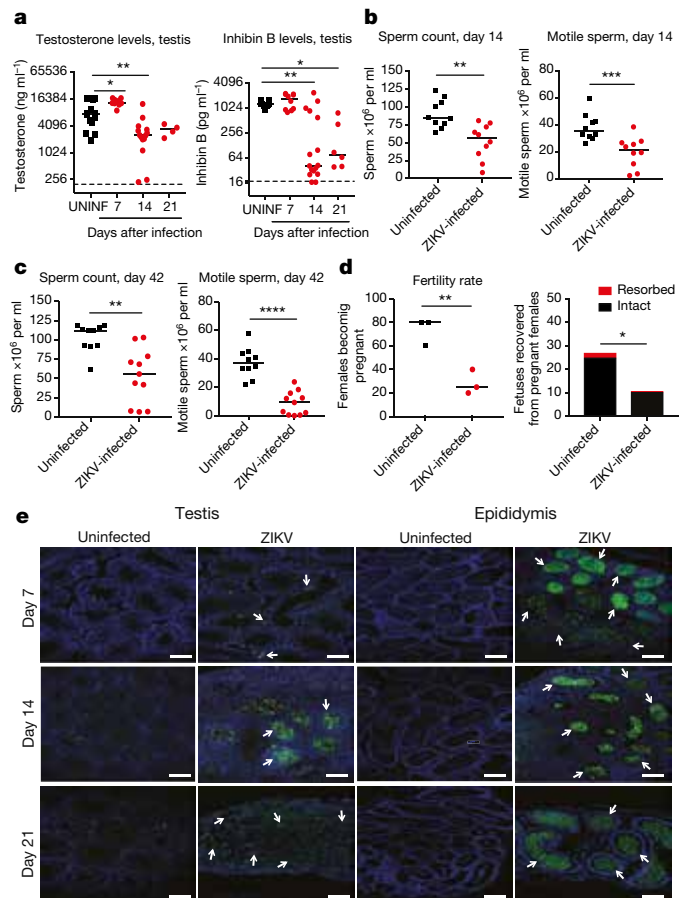


Figure 4 | Consequences of ZIKV infection of the testis and epididymis.

a, Testosterone (left) and inhibin B (right) levels of testis homogenates from uninfected (UNINF) and ZIKV-infected (days 7, 14 or 21) mice. **b**, **c**, Computer-assisted sperm analysis (total (left) and motile (right)) on samples obtained from the cauda epididymis of ZIKV-infected males immediately after euthanasia at days 14 (**b**) or approximately 42 (41–48 days) (**c**) after infection or age-matched uninfected males. **a**–**c**, Results are pooled from 2–5 independent experiments. Bars indicate median values, and differences between uninfected and ZIKV-infected animals were evaluated (**a**; **P* < 0.05; ***P* < 0.01; ****P* < 0.001; *****P* < 0.0001; ANOVA (Kruskal-Wallis) with a multiple comparison correction; **b**, **c**; ***P* < 0.01; ****P* < 0.001; *****P* < 0.0001; Mann-Whitney test). Dashed lines indicate the limit of sensitivity of the assay. **d**, Fertility studies. Age-matched uninfected or ZIKV-infected males (at days 7, 16 or 26 after infection (*n* = 4–5 male mice for each time point)) were mated with individual 8-week-old female wild-type mice (*n* = 4–5 females per round with 3 independent rounds performed) for five days and then separated. Ten days later, we evaluated the pregnancy rate (left, symbols correspond to the percentage of five females becoming pregnant for a given trial) and the total number of viable or resorbed fetuses for each round (right) (**P* < 0.05; ***P* < 0.01; unpaired Student's *t*-test). **e**, TUNEL staining of testis (left) and epididymis (right) from uninfected or ZIKV-infected mice at days 7, 14 or 21. TUNEL staining in germ and somatic cells of the testis and luminal sperm in the epididymis is shown (green staining and white arrows). The images are representative of several independent experiments. Scale bars, 100 μ m.

for spermatogenesis and express high levels of the TAM receptors Tyro3, Axl and Mertk¹¹. Because Axl has recently been postulated as an entry factor for ZIKV infection into cells^{12–16}, we assessed the effect of a genetic deficiency of Axl on ZIKV infection of the testis and epididymis. As we found high levels of infection in the testis and epididymis in *Axl*^{−/−} mice (Extended Data Fig. 4a), this TAM receptor probably does not have an essential role in ZIKV pathogenesis in the male reproductive tract. ISH showed strong staining of viral RNA in both Sertoli and germ cells in *Axl*^{−/−} mice at day 7 after ZIKV infection (Extended Data Fig. 4b).

The histological analysis showed that injury of the testis was associated with inflammatory cell infiltration. To assess the role of adaptive immune cells in the pathogenesis of acute disease, we inoculated *Rag1*^{-/-} mice, which lack both mature B and T cells, with ZIKV after a similar treatment with anti-Ifnar1 antibody. At day 7, we observed high levels of viral RNA in all male reproductive tract tissues (Extended Data Fig. 4a). At day 13, we observed ZIKV RNA in germ and Sertoli cells in *Rag1*^{-/-} mice, and this was associated with a decrease in TRA98⁺ germ cells and Lin28a⁺ spermatogonia and breakdown of the BTB. However, interstitial Leydig cells remained in ZIKV-infected *Rag1*^{-/-} mice even though the architecture of the seminiferous tubules was altered (Extended Data Fig. 4c–d). Thus, damage to the testis appears to be mediated both by ZIKV infection and adaptive immune responses.

To determine the functional consequences of ZIKV-Dakar infection in the testis, we measured the levels of two hormones important for spermatogenesis, testosterone and inhibin B, which are produced by Leydig and Sertoli cells, respectively. At day 7 after ZIKV infection, testosterone levels in homogenates of testes were increased, possibly because of the altered cellular physiology or inflammatory environment associated with viral replication¹⁷. By day 14, testosterone levels in ZIKV-infected mice were decreased and remained low at 21 days (Fig. 4a, left). Inhibin B levels were also reduced in ZIKV-infected testes at days 14 and 21 after infection (Fig. 4a, right). We observed diminished total and motile sperm counts from fluid collected from the cauda epididymis at 14 (Fig. 4b) or approximately 42 (Fig. 4c) days after ZIKV inoculation, which was consistent with extensive damage to the seminiferous tubules (Fig. 2d–f and Extended Data Figs 2, 5a, b). We also observed reduced rates of pregnancy and numbers of viable fetuses from females mated with ZIKV-infected males compared to uninfected males (Fig. 4d). Consistent with substantial injury to the testis, there was marked cell death in the seminiferous tubules and lumen of the epididymis at multiple time points, as indicated by TUNEL staining (Fig. 4e) and loss of cellularity (Fig. 3e, f). Thus, in mice, the injury to the male reproductive tract due to ZIKV infection results in decreased sex hormone production and oligospermia. ZIKV pathogenesis in the testis appears distinct from that of mumps virus, which preferentially infects interstitial Leydig cells and causes highly inflammatory acute orchitis^{18,19}.

In most human infections, ZIKV causes a mild febrile illness associated with rash and conjunctivitis. However, severe phenotypes are now appreciated, including Guillain-Barré syndrome^{20,21} and congenital abnormalities in fetuses²². ZIKV can be transmitted sexually, in contrast to related flaviviruses, as infectious virus persists in the semen of males^{23–25} for up to 80 days after symptom onset². Our experiments with mouse-adapted ZIKV-Dakar show that infection causes testicular and epididymal damage in mice that can progress to reductions in key sex hormones, destruction of germ and somatic cells in the testis, and loss of mature sperm and fertility. Sertoli cells may be a key target for ZIKV in the testis, resulting in cell dysfunction, detachment from the basement membrane and dissolution of the BTB. Infiltrating inflammatory cells may amplify destruction of the testicular architecture. Although further studies are required, this pathologic process results in decreased male fertility, at least in mice. While Axl is not required for infection of the mouse testis, other TAM or T-cell immunoglobulin and mucin domain (TIM)¹⁵ receptors could be important for ZIKV tropism.

The establishment of a model of male reproductive tract injury after ZIKV infection will allow the rapid testing of new classes of therapeutic agents^{26,27} or vaccines²⁸ to mitigate or prevent disease. Although our data are concerning for yet another unanticipated clinical manifestation of ZIKV infection, we acknowledge these results reflect studies exclusively performed in mice. Nonetheless, genitourinary signs and symptoms, including haematospermia, dysuria and perineal pain^{5,6,29}, have been reported in ZIKV-infected humans and ZIKV was recently

detected in human spermatozoa³⁰. Longitudinal studies monitoring ZIKV infection in semen and sperm counts seem warranted to define the extent and consequences of this disease process in affected human males.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 September; accepted 24 October 2016.

Published online 31 October 2016.

- Coyne, C. B. & Lazear, H. M. Zika virus—reigniting the TORCH. *Nat. Rev. Microbiol.* **14**, 707–715 (2016).
- Matheron, S. et al. Long-lasting persistence of Zika virus in semen. *Clin. Infect. Dis.* **63**, 1264 (2016).
- Lazear, H. M. et al. A mouse model of Zika virus pathogenesis. *Cell Host Microbe* **19**, 720–730 (2016).
- Rossi, S. L. et al. Characterization of a novel murine model to study Zika virus. *Am. J. Trop. Med. Hyg.* **94**, 1362–1369 (2016).
- Foy, B. D. et al. Probable non-vector-borne transmission of Zika virus, Colorado, USA. *Emerg. Infect. Dis.* **17**, 880–882 (2011).
- Musso, D. et al. Potential sexual transmission of Zika virus. *Emerg. Infect. Dis.* **21**, 359–361 (2015).
- Deckard, D. T. et al. Male-to-Male Sexual Transmission of Zika Virus—Texas, January 2016. *MMWR Morb. Mortal. Wkly. Rep.* **65**, 372–374 (2016).
- Grant, A. et al. Zika virus targets human STAT2 to inhibit type I Interferon signaling. *Cell Host Microbe* **19**, 882–890 (2016).
- Morrow, C. M. et al. ETV5 is required for continuous spermatogenesis in adult mice and may mediate blood testes barrier function and testicular immune privilege. *Ann. NY Acad. Sci.* **1120**, 144–151 (2007).
- Mullen, T. E. Jr, Kiessling, R. L. & Kiessling, A. A. Tissue-specific populations of leukocytes in semen-producing organs of the normal, hemicastrated, and vasectomized mouse. *AIDS Res. Hum. Retroviruses* **19**, 235–243 (2003).
- Lu, Q. et al. Tyro-3 family receptors are essential regulators of mammalian spermatogenesis. *Nature* **398**, 723–728 (1999).
- Savidis, G. et al. Identification of Zika virus and dengue virus dependency factors using functional genomics. *Cell Reports* **16**, 232–246 (2016).
- Nowakowski, T. J. et al. Expression analysis highlights AXL as a candidate Zika virus entry receptor in neural stem cells. *Cell Stem Cell* **18**, 591–596 (2016).
- Hamel, R. et al. Biology of Zika virus infection in human skin cells. *J. Virol.* **89**, 8880–8896 (2015).
- Tabata, T. et al. Zika virus targets different primary human placental cells, suggesting two routes for vertical transmission. *Cell Host Microbe* **20**, 155–166 (2016).
- Liu, S., DeLalio, L. J., Isakson, B. E. & Wang, T. T. AXL-Mediated productive infection of human endothelial cells by Zika virus. *Circ. Res.* CIRCRESAHA.116.309866 (2016).
- Dierich, A. et al. Impairing follicle-stimulating hormone (FSH) signaling in vivo: targeted disruption of the FSH receptor leads to aberrant gametogenesis and hormonal imbalance. *Proc. Natl Acad. Sci. USA* **95**, 13612–13617 (1998).
- Wu, H. et al. Mumps virus-induced innate immune responses in mouse Sertoli and Leydig cells. *Sci. Rep.* **6**, 19507 (2016).
- Le Goffic, R. et al. Mumps virus decreases testosterone production and gamma interferon-induced protein 10 secretion by human Leydig cells. *J. Virol.* **77**, 3297–3300 (2003).
- Oehler, E. et al. Zika virus infection complicated by Guillain-Barre syndrome—case report, French Polynesia, December 2013. *Euro Surveill.* **19**, 20720 (2014).
- Carteaux, G. et al. Zika virus associated with Meningoencephalitis. *N. Engl. J. Med.* **374**, 1595–1596 (2016).
- Brasil, P. et al. Zika virus infection in pregnant women in Rio de Janeiro—preliminary report. *N. Engl. J. Med.* 10.1056/NEJMoa1602412 (2016).
- Mansuy, J. M. et al. Zika virus in semen of a patient returning from a non-epidemic area. *Lancet Infect. Dis.* **16**, 894–895 (2016).
- Turmel, J. M. et al. Late sexual transmission of Zika virus related to persistence in the semen. *Lancet* **387**, 2501 (2016).
- D'Ortenzio, E. et al. Evidence of sexual transmission of Zika virus. *N. Engl. J. Med.* **374**, 2195–2198 (2016).
- Zhao, H. et al. Structural basis of Zika Virus-specific antibody protection. *Cell* **166**, 1016–1027 (2016).
- Barrows, N. J. et al. A Screen of FDA-Approved drugs for inhibitors of Zika virus infection. *Cell Host Microbe* **20**, 259–270 (2016).
- Larocca, R. A. et al. Vaccine protection against Zika virus from Brazil. *Nature* **536**, 474–478 (2016).
- Torres, J. R., Martinez, N. & Moros, Z. Microhematospermia in acute Zika virus infection. *Int J Infect Dis.* **51**, 127 (2016).
- Mansuy, J. M. et al. Zika virus in semen and spermatozoa. *Lancet Infect. Dis.* **16**, 1106–1107 (2016).

Acknowledgements NIH grants (R01 AI073755 and R01 AI104972 to M.S.D., R01 HD065435 and R01HD083895 to K.H.M., and T32 AI007163 (E.F.)) supported this work. This work was supported by the Washington University Institute of Clinical and Translational Sciences (UL1 TR000448 from the National Center for Advancing Translational Sciences and P41 GM103422-35 from the National Institute of General Medical Sciences to K.H.M.), as well as a grant from the Veteran Affairs Office of Research and Development IO1BX007080 to K.H.M.). The authors thank J. Miner, T. Pierson, P. A. Felder and J. Halabi for technical assistance, manuscript review and data analysis. The testosterone and inhibin B assays were processed by the University of Virginia Center for Research in Reproduction Ligand Assay and Analysis Core, which is supported by the Eunice Kennedy Shriver NICHD/NIH (NCTRI) Grant P50-HD28934.

Author Contributions J.G., P.E., S.M.S., E.F., A.D., D.J.P., J.M.R., E.A.C. and V.S. performed the experiments. M.J.G. provided key reagents. J.G., P.E., S.M.S. and E.F. performed data analysis. M.S.D., P.E. and K.H.M. wrote the initial draft of the manuscript, with all other authors contributing to editing into the final form.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.H.M. (moleyk@wustl.edu) or M.S.D. (diamond@wusm.wustl.edu).

METHODS

Ethics statement. This study was carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The protocols were approved by the Institutional Animal Care and Use Committee at the Washington University School of Medicine (Assurance number A3381-01). Inoculations were performed under anaesthesia induced and maintained with ketamine hydrochloride and xylazine, and all efforts were made to minimize animal suffering.

Viruses. ZIKV strain H/PF/2013 (French Polynesia, 2013) was provided by the Arbovirus Branch of the Centers for Disease Control and Prevention with permission (X. de Lamballerie). ZIKV strain Dakar 41519 (Senegal, 1984) was provided by the World Reference Center for Emerging Viruses and Arboviruses (R. Tesh) and passaged twice in *RagI*^{-/-} mice to create a mouse-adapted, more pathogenic variant of ZIKV-Dakar (M. Gorman and M. Diamond, unpublished results). DENV-2 (strain D2S20) was obtained as a gift (S. Shrestha). Virus stocks were propagated in mycoplasma-free Vero cells (ATCC) and titrated by focus-forming assay (FFA), as described previously³.

Mouse infection experiments. Wild-type C57BL/6 mice were purchased commercially (Jackson Laboratories) and congenic *RagI*^{-/-} mice were bred at Washington University in a pathogen-free facility. Congenic *Axl*^{-/-} mice were described previously³¹. Seven-week-old mice were inoculated by subcutaneous route in the footpad with 10³ (H/PF/2013) or 10⁶ (Dakar 41519 or DENV-2) FFU in a volume of 50 µl. One-day before inoculation with virus, mice were treated with 0.5 or 2 mg of an Ifnar1-blocking mouse antibody (MAR1-5A3) or isotype control mouse antibody (GIR-208) by intraperitoneal injection³. At different days after infection, tissues were collected and processed as described below. Testis and epididymis collected from infected male mice were processed for haematoxylin and eosin staining, immunofluorescence and confocal microscopy, ISH and viral titer analysis as described previously³². Testes were also examined macroscopically and weighed. At days 14, 21 and around 42 after ZIKV-Dakar infection, the macroscopic damage as indicated by a reduction in size was often uniformly bilateral, although some asymmetry in testis (right versus left) size was observed. Randomization and blinding of the animal experiments were not performed, and sample sizes were not calculated beforehand.

Computer-assisted sperm analysis (CASA). Mature sperm from the cauda epididymis of uninfected or virus-infected mice were collected immediately after euthanasia as reported earlier³³. The sperm suspension in vitrofret medium (Cook Medical) was analysed for total sperm count by CASA using the HTM-IVOS Vs12 integrated visual optical system motility analyser (Hamilton-Thorne Research) as described previously³⁴. For studies at around 42 day after infection, mice were killed at day 41 (*n* = 3), 42 (*n* = 4), 43 (*n* = 3), and 48 (*n* = 1) after infection with 10³ to 10⁶ FFU of ZIKV-Dakar. All measurements of total and motile sperm were made within 60 min of dissection of the cauda epididymis.

Testosterone and inhibin B levels. Total homogenates of testes from uninfected or ZIKV-infected mice were assayed for testosterone and inhibin B levels by radioimmunoassay as described¹⁷ using the Research in Reproduction Ligand Assay and Analysis Core at the University of Virginia.

Fertility studies. Age-matched uninfected or ZIKV-infected wild-type C57BL/6 males (at days 7, 16 or 26 after infection, *n* = 4–5 at each time point) were mated with single 8-week-old female wild-type C57BL/6 mice. Five days later, males were removed from the cage to isolate the females. Ten days later, female mice (*n* = 14–15 for each group) were euthanized and evaluated for pregnancy, and the number of viable or resorbed fetuses was counted. Because sperm from mice can be obtained only at euthanasia, we were unable to perform longitudinal studies and directly correlate sperm counts after ZIKV infection with fertility rates.

Viral burden. ZIKV- or DENV-infected mice were euthanized on specific days. Testis, epididymis and other tissues were weighed and homogenized with zirconia beads in a MagNA Lyser instrument (Roche Life Science) in 200 µl PBS.

All homogenized tissues from infected animals were stored at –80 °C. With some samples, viral burden was determined by plaque assay on Vero cells³⁵. Sperm were subjected to three rapid freeze-thaw cycles to release infectious virus. Other samples were extracted with the RNeasy Mini Kit. ZIKV and DENV RNA levels were determined by one-step quantitative reverse transcriptase PCR (qRT-PCR) on an ABI 7500 Fast Instrument using standard cycling conditions. Viral burden was expressed on a log₁₀ scale as viral RNA equivalents per g or ml after comparison with a standard curve produced using serial tenfold dilutions of ZIKV or DENV RNA as described previously³⁵. For ZIKV, the following primer sets were used: 1183F: 5'-CCACCAATGTTCTCTTGCAGACATATTG-3'; 1268R: 5'-TTCGGA CAGCCGTTGTCCAACACAAG-3'; and probes (1213F): 5'-56-FAM/AGCCTA CCT TGACAAGCAGTC/3IABkFQ-3'.

Histology and immunohistochemistry. Tissues were collected after death and fixed overnight in 4% paraformaldehyde (PFA) in PBS. Subsequently, 5-µm-thick testis and epididymal sections from infected and uninfected mice were processed for histology by haematoxylin and eosin staining. For immunohistochemistry, the tissue sections were incubated with mouse primary monoclonal anti-CD45 (610266; BD Biosciences), anti-ETV5 (ab102010; Abcam), anti-GATA4 (ab84593; Abcam), rabbit polyclonal anti-Lin28a (3978S, Cell Signaling), rat polyclonal anti-TRA98 (ab82527; Abcam), rat polyclonal anti-F4/80 (ab6640; Abcam), or goat polyclonal anti-β3-HSD antibodies (SC-30820, Santa Cruz Biotechnology). After washing, slides were stained with Alexa Fluor 488- or, Alexa Fluor 546-conjugated goat anti-rabbit, goat anti-mouse or donkey anti-goat (1:1,000; A11008, A11081, A11030 or A11056; ThermoFisher Scientific) secondary antibodies for 1 h, and mounted with prolong gold anti-fade mount containing the nuclear counter stain, DAPI (ThermoFisher Scientific). Immunostaining was detected by confocal microscopy (Leica SPE100, Germany).

Viral RNA *in situ* hybridization. RNA ISH was performed using RNAscope 2.5 (Advanced Cell Diagnostics) according to the manufacturer's instructions. PFA-fixed paraffin-embedded tissue sections were deparaffinized by incubating for 60 min at 60 °C. Endogenous peroxidases were quenched with H₂O₂ for 10 min at room temperature. Slides were boiled for 15 min in RNAscope Target Retrieval Reagents and incubated for 30 min in RNAscope Protease Plus before probe hybridization. The probe targeting ZIKV RNA was designed and synthesized by Advanced Cell Diagnostics (Catalog #467871). Positive (targeting *plp2a* gene) and negative (targeting bacterial gene *dapB*) control probes also were obtained from Advanced Cell Diagnostics (Catalog #312471 and #310043, respectively). Tissues were counterstained with Gill's haematoxylin and visualized using bright-field microscopy.

Data analysis. All data were analysed with GraphPad Prism software. For viral burden analysis, the log₁₀ transformed titres were analysed by the Mann–Whitney test or a Kruskal–Wallis one-way ANOVA. A *P* value of <0.05 indicated statistically significant differences.

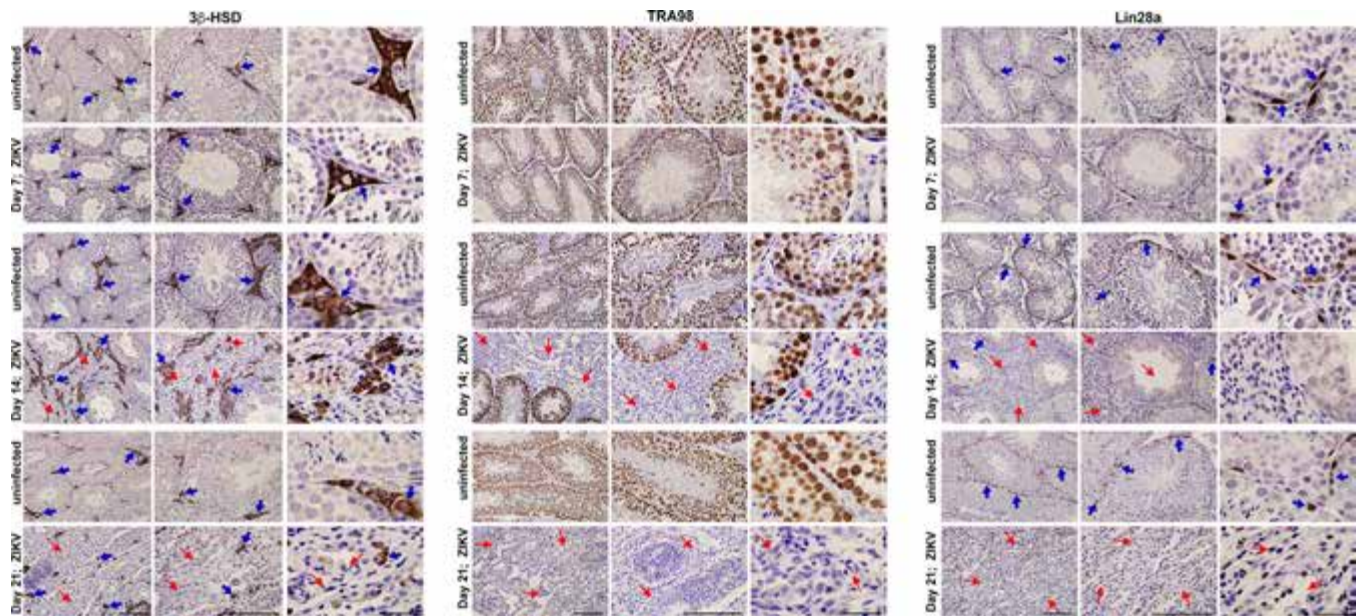
Data availability. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

- Lu, Q. & Lemke, G. Homeostatic regulation of the immune system by receptor tyrosine kinases of the Tyro 3 family. *Science* **293**, 306–311 (2001).
- Esakky, P., Hansen, D. A., Drury, A. M. & Moley, K. H. Molecular analysis of cell type-specific gene expression profile during mouse spermatogenesis by laser microdissection and qRT-PCR. *Reprod. Sci.* **20**, 238–252 (2013).
- Hansen, D. A., Esakky, P., Drury, A., Lamb, L. & Moley, K. H. The aryl hydrocarbon receptor is important for proper seminiferous tubule architecture and sperm development in mice. *Biol. Reprod.* **90**, 8 (2014).
- Goodson, S. G., Zhang, Z., Tsuruta, J. K., Wang, W. & O'Brien, D. A. Classification of mouse sperm motility patterns using an automated multiclass support vector machines model. *Biol. Reprod.* **84**, 1207–1215 (2011).
- Miner, J. J. *et al.* Zika virus infection during pregnancy in mice causes placental damage and fetal demise. *Cell* **165**, 1081–1091 (2016).



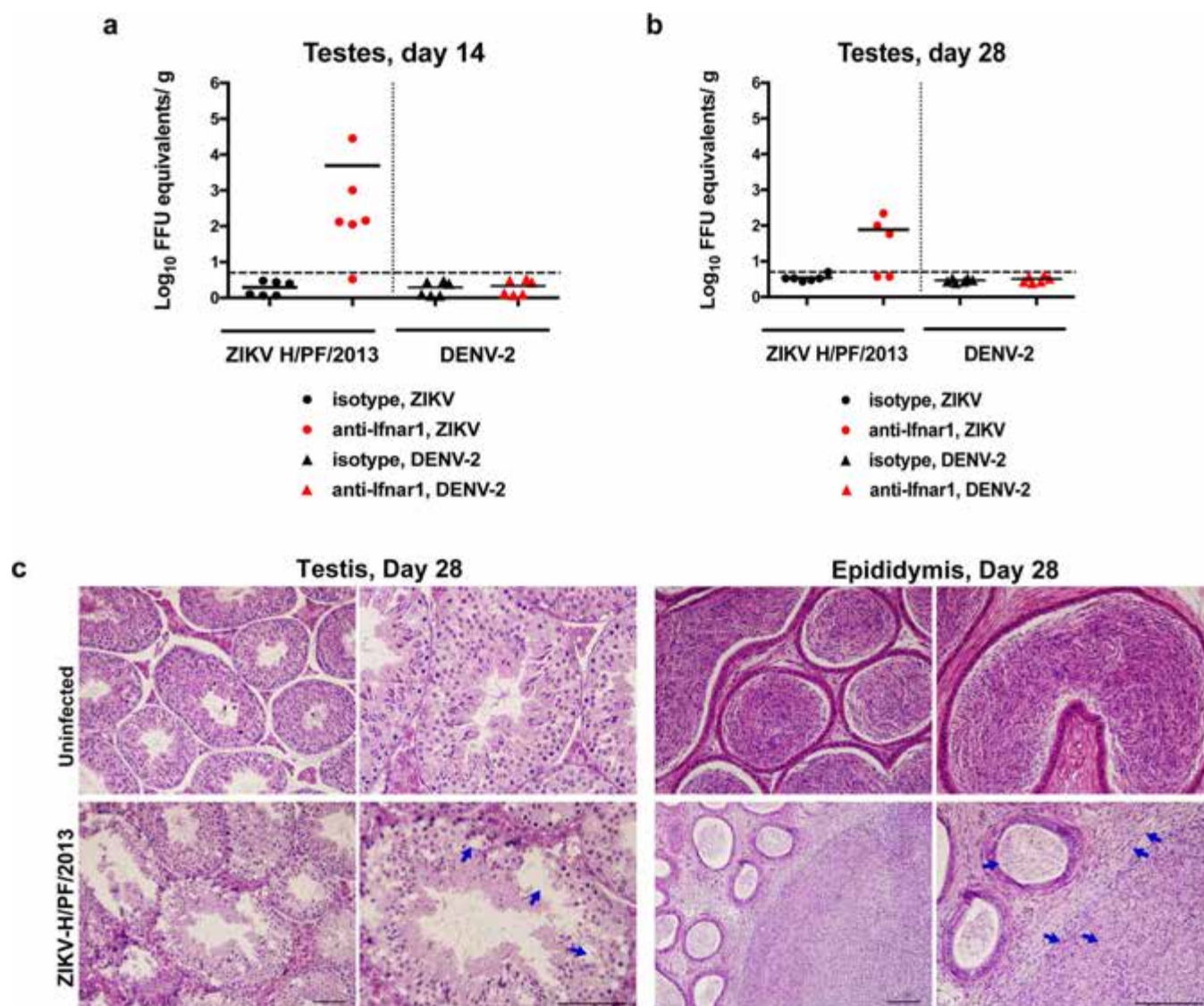
Extended Data Figure 1 | ZIKV infection of mature sperm. Mature sperm was collected from the cauda epididymis of uninfected (left) or ZIKV-infected (day 7, right) mice and processed by ISH with a ZIKV-specific probe. Staining for viral RNA is seen in the ZIKV-infected samples at the head (inset, red arrow) and in the cytoplasmic droplets

(green arrows) in the sperm flagellum. Scale bar, 50 μm . Staining was quantified by microscopy: uninfected: 81 sperm counted, 0 positive for staining in head, 0 positive for staining in tail; and ZIKV-infected: 93 sperm counted, 25 (27%) positive for staining in head, 57 (61%) positive for staining in tail.



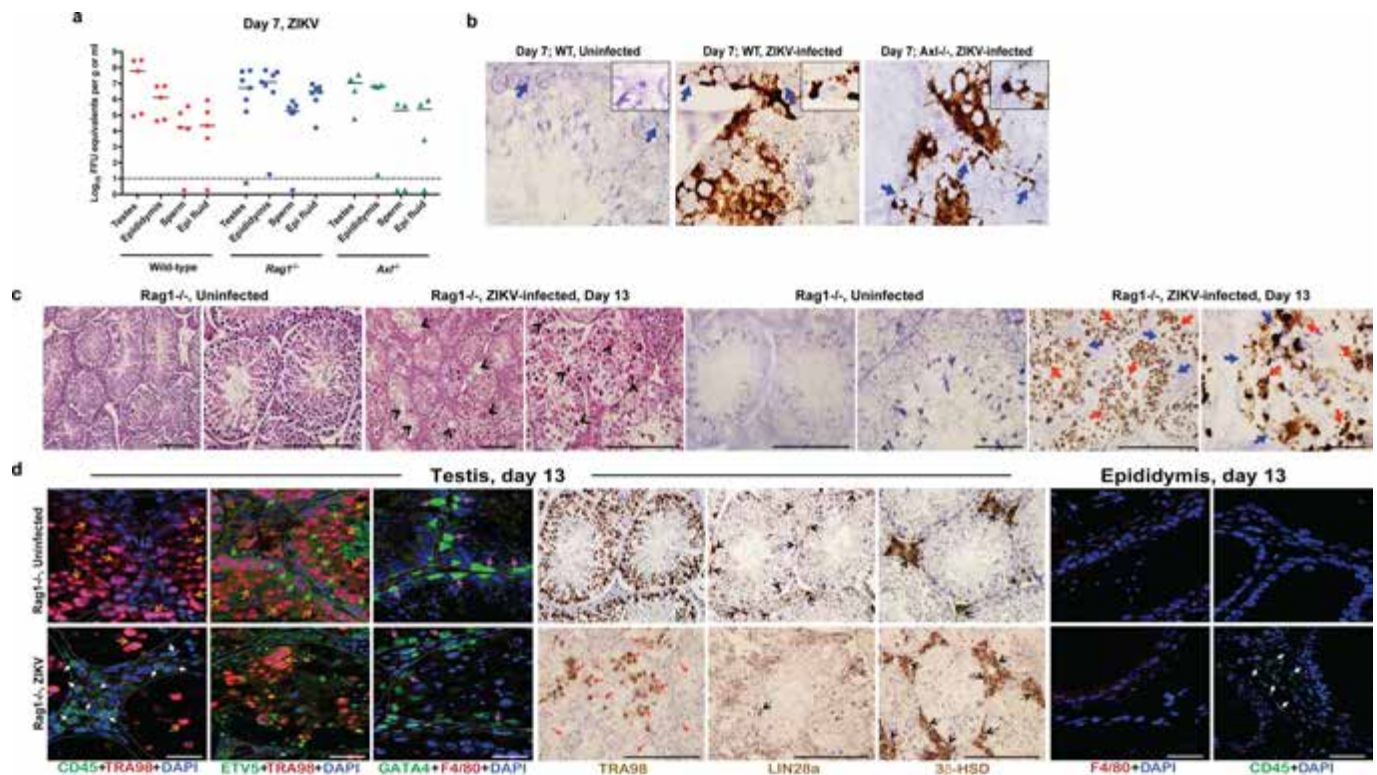
Extended Data Figure 2 | Temporal loss of cellularity in the testis after ZIKV infection. Seven-week-old wild-type C57BL/6 mice were treated with 0.5 mg of anti-Ifnar1 at day -1 before subcutaneous inoculation of mouse-adapted ZIKV-Dakar. Immunohistochemical analysis was performed on testis tissue collected from uninfected (top) or ZIKV-infected animals (days 7, 14 or 21 after infection; bottom) at $20\times$ (left), $40\times$ (middle) and $100\times$ (right image) magnification. Staining was

performed with antibodies against 3β -HSD (Leydig cells, top), TRA98 (germ cells, middle), and Lin28a (type A undifferentiated and type B spermatogonia, bottom). Blue arrows indicate staining of Leydig cells (top) and spermatogonial stem cells (bottom). Red arrows indicate areas of virus-induced damage and loss of tissue integrity and specific cellularity. Scale bars, 200, 200 and $50\mu\text{m}$ for the grouping of the three sets of images.



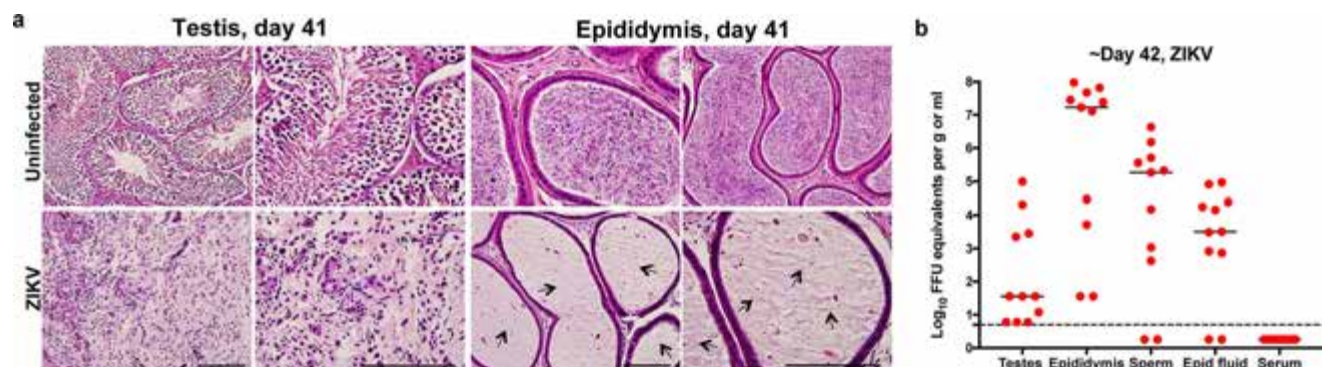
Extended Data Figure 3 | Histology of the testis at day 28 after infection with ZIKV-H/PF/2013. **a, b,** Seven-week-old wild-type C57BL/6 mice were treated with PBS or anti-Ifnar1 at day -1 before subcutaneous inoculation in the footpad with 10^3 FFU of ZIKV-H/PF/2013 or 10^6 FFU of DENV-2. Testes were collected at day 14 (**a**) or 28 (**b**) after infection and analysed for viral RNA by qRT-PCR. Results are pooled from two independent biological experiments and each symbol represents data from an individual mouse. Bars indicate mean values. **c,** Histological analysis

of paraformaldehyde-fixed testis (left) and epididymis (right) collected from uninfected or ZIKV-infected animals at day 28 at $20\times$ (left) and $40\times$ (right) magnification. Arrows indicate loss of germ cells and vacuoles in the testis, involution of epididymal lumens with a mass of residual sperm, and thickened epithelium. The images are representative of several independent experiments. Scale bars ($50\ \mu\text{m}$ (left panel of each tissue) and $200\ \mu\text{m}$ (right panel of each tissue)) are indicated in the bottom right corner of the panels.



Extended Data Figure 4 | ZIKV infection of the testis and epididymis at day 7 in *Axl*^{-/-} and *Rag1*^{-/-} mice. Seven-week-old wild-type (WT), *Axl*^{-/-} or *Rag1*^{-/-} C57BL/6 mice were treated with 0.5 mg of anti-Ifnar1 at day -1 before subcutaneous inoculation in the footpad with 10⁶ FFU of mouse-adapted ZIKV-Dakar. **a**, The indicated tissues were collected at day 7 after infection and analysed for viral RNA by qRT-PCR. Each symbol corresponds to data from an individual mouse and was produced from at least two independent experiments. Dashed lines indicate limit of detection of the assays. **b**, ISH of testis from uninfected or ZIKV-infected *Rag1*^{-/-} mice at day 7 with a ZIKV-specific probe. Dark blue arrows indicate Sertoli cells. Inset, in sections from infected wild-type and *Axl*^{-/-} mice, the cytoplasm of Sertoli cells is positive for ZIKV RNA (dark brown) with signal absent from prominent nuclei and nucleoli. Scale bar, 50 μm. **c**, Histology (haematoxylin and eosin, left two) and ISH (right two) of testis from age-matched uninfected or ZIKV-infected (day 13) *Rag1*^{-/-} mice at 20× (left) and 40× (right) magnification for each pair. Scale bars, 200 (the second, fourth, fifth and seventh image from the left) and 50 μm (the first, third, sixth and eighth image from the left).

In haematoxylin and eosin-stained testis sections, arrows indicate loss of germ cells and presence of multi-nucleated giant and necrotic cells from ZIKV-infected *Rag1*^{-/-} mice. In ISH, red and blue arrows indicate distribution of ZIKV RNA and Sertoli cells, respectively. **d**, Immunofluorescence (three left and two right) and immunohistochemistry (three middle) staining of uninfected or ZIKV-infected (day 13) testes and epididymis from *Rag1*^{-/-} mice with antibodies to CD45, TRA98, ETV5, GATA4, LIN28a, 3β-HSD or F4/80 as described in Fig. 1 and Extended Data Fig. 2. Coloured arrows indicate staining for leukocytes (CD45, white), germ cells (TRA98, orange), Sertoli cells (GATA4, magenta), BTB (ETV5, green), type A undifferentiated and type B spermatogonia (Lin28a, black) and Leydig cells (3β-HSD, black). In the immunohistochemistry staining panels with TRA98, red arrows indicate dying or dead germ cells and tubules without germ cells. White lines demarcate tubules in the seminiferous epithelium. Scale bars, 200 μm for immunohistochemistry (middle three columns, for TRA98, LIN28a and 3β-HSD staining) and 50 μm for immunofluorescence (left three and right two columns). The images are representative of several different animals.



Extended Data Figure 5 | ZIKV infection of the testis and epididymis around day 42. **a**, Seven-week-old wild-type C57BL/6 mice were treated with anti-Ifnar1 at day -1 before subcutaneous inoculation in the footpad with 10^6 FFU of mouse-adapted ZIKV-Dakar. **a**, Testis (left) and epididymis (right) were collected at day 41 after infection or from age-matched uninfected mice, fixed with paraformaldehyde, sectioned, stained with haematoxylin and eosin, and imaged at a magnification of $20\times$ (left) and $40\times$ (right). Arrows show epididymal lumen void of sperm. The images are representative of sections from several independent

animals. Scale bars are indicated in the bottom right corner of the panels. Scale bars, $200\ \mu\text{m}$ (right columns for both testis and epididymis) and $50\ \mu\text{m}$ (left columns for both testis and epididymis). **b**, The indicated tissues and cells were collected around day 42 after infection (days 41 ($n=3$), 42 ($n=4$), 43 ($n=3$), and 48 ($n=1$)) and analysed for viral RNA by qRT-PCR. Dashed line indicates the limit of detection of the assay. Results are pooled from 2–3 independent biological experiments and each symbol represents data from an individual mouse. Bars indicate median values.

Neutralizing human antibodies prevent Zika virus replication and fetal disease in mice

Gopal Sapparapu^{1,2*}, Estefania Fernandez^{3*}, Nurgun Kose², Bin Cao⁴, Julie M. Fox⁵, Robin G. Bombardi², Haiyan Zhao³, Christopher A. Nelson³, Aubrey L. Bryan⁶, Trevor Barnes⁶, Edgar Davidson⁶, Indira U. Mysorekar^{3,4}, Daved H. Fremont³, Benjamin J. Doranz⁶, Michael S. Diamond^{3,5,7,8} & James E. Crowe Jr^{1,2,9}

Zika virus (ZIKV) is an emerging mosquito-transmitted flavivirus that can cause severe disease, including congenital birth defects during pregnancy¹. To develop candidate therapeutic agents against ZIKV, we isolated a panel of human monoclonal antibodies from subjects that were previously infected with ZIKV. We show that a subset of antibodies recognize diverse epitopes on the envelope (E) protein and exhibit potent neutralizing activity. One of the most inhibitory antibodies, ZIKV-117, broadly neutralized infection of ZIKV strains corresponding to African and Asian-American lineages. Epitope mapping studies revealed that ZIKV-117 recognized a unique quaternary epitope on the E protein dimer-dimer interface. We evaluated the therapeutic efficacy of ZIKV-117 in pregnant and non-pregnant mice. Monoclonal antibody treatment markedly reduced tissue pathology, placental and fetal infection, and mortality in mice. Thus, neutralizing human antibodies can protect against maternal-fetal transmission, infection and disease, and reveal important determinants for structure-based rational vaccine design efforts.

Recent ZIKV epidemics are linked to Guillain-Barré syndrome in adults and microcephaly in fetuses and newborn infants^{2–5}. Although ZIKV infection can potentially cause severe disease, specific treatments and vaccines for ZIKV are not currently available. We sought to isolate neutralizing human monoclonal antibodies (mAbs) with broad specificity against all ZIKV strains and protective activity *in vivo*. We tested the serological response of subjects who had previously been infected with ZIKV in diverse geographic locations. Serum from each subject contained antibodies that were shown by ELISA assays to react with ZIKV E protein and to neutralize infection of a contemporary Asian isolate (H/PF/2013) from French Polynesia (Fig. 1a, b). We studied the B cells of subject 1001 in greater detail. Based on the results of replicate assays, the frequency of B cells that secrete antibodies against ZIKV E protein in the peripheral blood was between 0.36% and 0.61% (Fig. 1c, d). We next tested the reactivity of antibodies with domain III (DIII) of the E protein from ZIKV or the related dengue (DENV) and West Nile (WNV) viruses. Only a subset (6%) of the ZIKV-E-reactive antibodies bound to DIII, and most were specific for ZIKV (Fig. 1c). Comparative binding to a wild-type or mutant ZIKV E protein lacking the conserved fusion loop epitope in DII (mutant denoted hereafter as E-FLM) established immunodominance (binding around 70% of mAbs) of the fusion loop.

We obtained 29 cloned hybridomas secreting mAbs that bound to ZIKV E protein from the cells of three donors (mAb ZIKV-195 from subject 1011, mAb ZIKV-204 from subject 973, and the remaining 27 mAbs from subject 1001). All of the mAbs except for one belonged to the IgG1 isotype (two could not be determined), with an equal

distribution of κ and λ light chains (Extended Data Table 1); sequence analysis of cDNA of the antibody variable gene regions revealed that each mAb represented an independent clone (Extended Data Table 1). We determined the half-maximal effective concentrations for binding

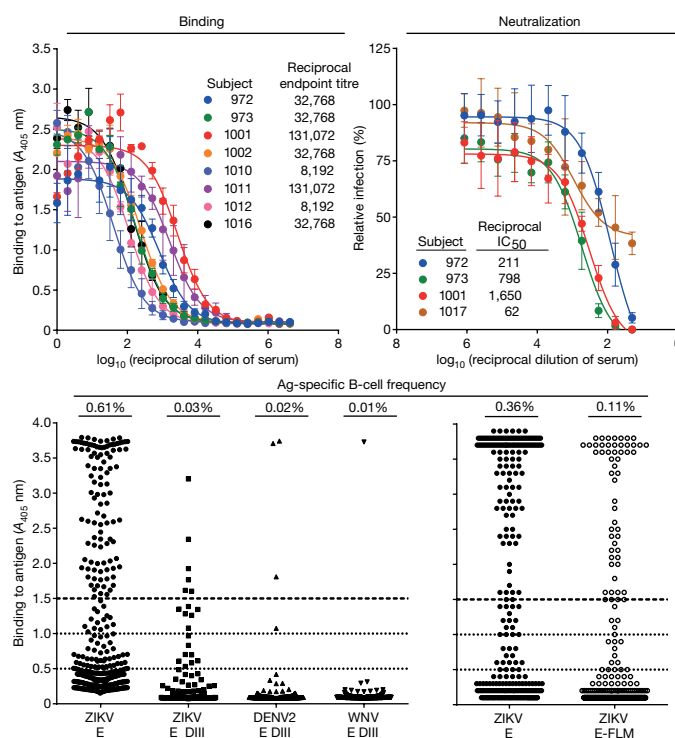


Figure 1 | Human antibody and B-cell response to ZIKV infection.

a, b, Serum samples from humans with a previous ZIKV infection were tested for binding to ZIKV E protein in ELISA (**a**) (with two technical replicates) and neutralization of ZIKV (**b**) (at least two independent repeats in triplicate). Subject 1001 had the highest endpoint titre in the binding assay and displayed potent neutralizing activity. Subject 657 was a control without history of exposure to ZIKV. **c**, Supernatants of Epstein-Barr virus (EBV)-transformed B-cell cultures from subject 1001 were tested for binding to ZIKV E or DIII of ZIKV E or related flavivirus E proteins; the WNV-reactive clone and all but one DENV-reactive B-cell line also reacted with ZIKV E protein. The frequency of antigen-specific cells against each viral protein was determined with a threshold absorbance value at 405 nm (A_{405} nm) of 1.5 as indicated. **d**, In four additional separate B-cell transformation experiments, the frequency of B cells reactive with intact ZIKV E or E-FLM was determined.

¹Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA. ²The Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

³Department of Pathology & Immunology, Washington University School of Medicine, St Louis, Missouri, USA. ⁴Department of Obstetrics and Gynecology, Washington University School of Medicine, St Louis, Missouri, USA. ⁵Department of Medicine, Washington University School of Medicine, St Louis, Missouri, USA. ⁶Integral Molecular, Philadelphia, Pennsylvania, USA.

⁷Department of Molecular Microbiology, Washington University School of Medicine, St Louis, Missouri, USA. ⁸Center for Human Immunology and Immunotherapy Programs, Washington University School of Medicine, St Louis, Missouri, USA. ⁹Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, Tennessee, USA.

*These authors contributed equally to this work.

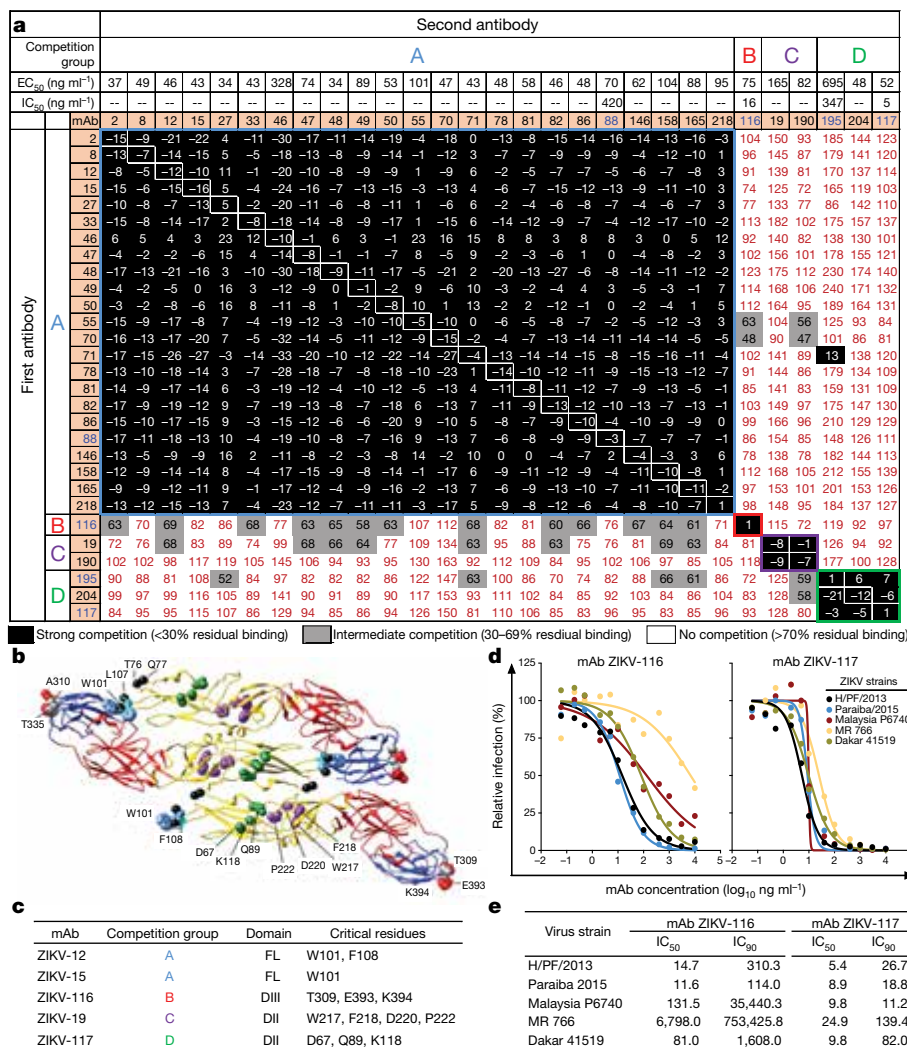


Figure 2 | Characterization of anti-ZIKV mAbs. **a**, We tested 29 mAbs in binding, neutralization, and competition binding assays. The EC₅₀ against ZIKV E and the IC₅₀ (by focus reduction neutralization test) against H/PF/2013 strain for neutralizing antibodies (highlighted in blue) are shown. The mAbs are displayed in four groups (A, B, C or D) based on a competition binding assay. The values are the percentage of binding that occurred during competition compared to non-competed binding, which was normalized to 100% and the range of competition is indicated by the box colours. Black filled boxes indicate strongly competing pairs (residual binding <30%), grey filled boxes indicate intermediate competition (residual binding 30–69%), and white filled boxes indicate non-competing pairs (residual binding ≥ 70%). The IC₅₀ against H/PF/2013 strain for neutralizing antibodies is shown with neutralizing clones highlighted in

to ZIKV E protein (EC₅₀) and neutralization (IC₅₀) of infection (Fig. 2a, Extended Data Fig. 1); most of the mAbs bound to E protein at low concentrations (EC₅₀ < 100 ng ml⁻¹), whereas only four of the 29 mAbs exhibited strong neutralizing activity (IC₅₀ = 5–420 ng ml⁻¹). We next determined how many antigenic sites on ZIKV E were recognized using quantitative competition binding. We identified four major competition groups (designated A, B, C or D). Group A mAbs had 23 members that were directed against the fusion loop in DII, as determined by differential binding to E and E-FLM (Extended Data Fig. 1), and had only one clone (ZIKV-88) with moderate neutralizing potency. The group B mAb ZIKV-116 neutralized ZIKV infection and bound to E, DIII and E-FLM. Group C mAbs (ZIKV-19 and ZIKV-190) bound to E and E-FLM weakly, but did not potently neutralize infection. The group D mAb ZIKV-195 neutralized with moderate potency and was similar in binding to both E and E-FLM. The most inhibitory group D mAb, ZIKV-117, bound to both E and E-FLM weakly.

blue. **b**, A ribbon diagram of three protomers of ZIKV E (DI in red, DII in yellow and DIII in blue) is shown with critical residues highlighted as spheres from epitope mapping experiments for representative antibodies in each of the competition binding groups. The colours of the critical residues correspond to the competition group designation as in **a**. The mutations in the E-FLM and DIII-LR mutants are indicated by black and silver spheres, respectively. **c**, Representative mAbs from each competition binding group are listed with the domains and residues critical for binding. FL, fusion loop. **d**, Two mAbs were tested for neutralization of five strains of ZIKV. The concentrations (ng ml⁻¹) at which 50% or 90% neutralization occurred are listed in **e**. The neutralization data are pooled from at least three independent experiments performed in triplicate.

We mapped the epitopes of representative mAbs using a shotgun alanine-scanning mutagenesis library⁶ of ZIKV prM and E protein variants (Fig. 2b, Extended Data Fig. 2). Loss-of-binding analysis confirmed that group A mAbs bound to the fusion loop in DII, whereas the group B mAb bound to DIII. Group B mAb ZIKV-116 bound to an epitope involving residues T309, E393 and K394 along the lateral ridge of DIII (DIII-LR), which was confirmed in an ELISA that showed reduced binding to DIII with mutations A310E and T335K in DIII-LR⁷. The epitope mapping studies suggest that the group D mAb ZIKV-117 binds specifically to DII across two adjacent dimers at the ‘dimer-dimer’ interface (Fig. 2c). We were unable to isolate virus neutralization escape mutant viruses for ZIKV-117, despite six passages in cell culture under mAb selection pressure.

Because of their potency, we assessed whether group B mAb ZIKV-116 and group D mAb ZIKV-117 could inhibit diverse ZIKV strains encompassing the African and Asian-American lineages. ZIKV-117

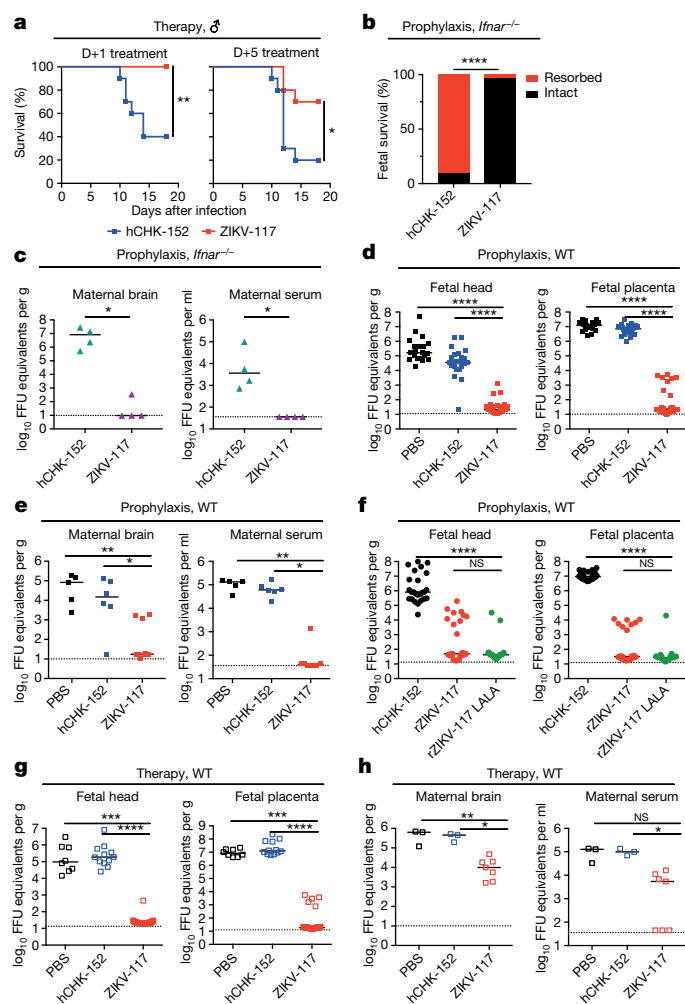


Figure 3 | Protective activity of ZIKV-117 in adult male and pregnant female mice. **a**, We treated 4–5-week-old wild-type male mice with 2 mg of anti-*Ifnar1* mAb followed by subcutaneous inoculation with 10^3 FFU of mouse-adapted ZIKV-Dakar. Mice were treated with a single 100 μ g or 250 μ g dose of isotype control mAb (hCHK-152) or ZIKV-117 on D+1 or D+5 ($n = 10$ per group from two independent experiments), respectively. Significance was analysed by the log-rank test (* $P < 0.05$; ** $P < 0.01$). **b**, *Ifnar1*^{-/-} female mice were mated with wild-type sires. At E5.5, dams were treated with 250 μ g of either hCHK-152 isotype control mAb or ZIKV-117. Bars indicate the median values and reflect data pooled from four independent experiments. Significance for fetal survival and viral RNA was analysed by chi-square (**b**; **** $P < 0.0001$) and Mann–Whitney (**c**; * $P < 0.05$) tests, respectively. **d–f**, Wild-type female mice were mated with wild-type sires. At E5.5, dams were treated with anti-*Ifnar1* mAb and one of the following: PBS (**d**, **e**), 250 μ g (**d–f**) of hCHK-152 isotype control mAb, 250 μ g of ZIKV-117 (**d–f**) or 250 μ g of ZIKV-117 LALA (**f**). At E6.5, dams were inoculated with 10^3 FFU of ZIKV-Dakar. **d–f**, Fetuses and placentas (**d**, **f**) and maternal brain and serum (**e**) were collected on E13.5 and viral RNA was measured by qRT–PCR. Bars indicate the median values of samples collected from three biological replicates (**d**, $n = 20–36$; **e**, $n = 5–9$; **f**, $n = 23–28$). Significance was analysed by ANOVA with a Dunn's multiple comparison test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$). **g**, **h**, Wild-type female mice were mated with wild-type sires. At E5.5, dams were treated with anti-*Ifnar1* mAb. At E6.5, dams were inoculated with 10^3 FFU of ZIKV-Dakar. At E7.5 (day +1 after infection), dams were treated with PBS, 250 μ g of hCHK-152 isotype control mAb, or 250 μ g of ZIKV-117. **g**, **h**, Fetuses and placentas (**g**) and maternal brain and serum (**h**) were collected on E13.5 and viral RNA was measured by qRT–PCR. Bars indicate the median values of samples collected from three biological replicates (**g**, $n = 8–20$; **h**, $n = 3–7$). Significance was analysed by ANOVA with Dunn's (**g**) or Tukey's (**h**) multiple comparisons test (* $P < 0.05$, *** $P < 0.001$, **** $P < 0.0001$). Dashed lines indicate the limit of detection of the assay.

neutralized all of the ZIKV strains tested, including two African (MR 766 and Dakar 41519), two Asian (Malaysia P6740 and H/PF/2013), and an American (Brazil Paraiba 2015) strain with IC₅₀ values of 5 to 25 ng ml⁻¹ (Fig. 2d, e). ZIKV-116 inhibited four of the five strains efficiently, but was inactive against MR 766, the original African strain (Fig. 2d, e). Alignment of the sequences of ZIKV H/PF/2013 and MR 766, with respect to residues in DIII-LR⁷ that ZIKV-116 binds, revealed only one difference (a conservative E393D change). Given these data, we hypothesize that the DIII-LR epitope of ZIKV-116 is displayed differently on MR 766 owing to allosteric effects of changes in other parts of the E protein, which could regulate epitope accessibility^{8,9}.

As recent studies have suggested that cross-reactive ZIKV-specific human mAbs can enhance DENV infection *in vivo*¹⁰, we tested whether these two ZIKV-neutralizing mAbs could bind to DENV-infected cells. ZIKV-117 showed a restricted type-specific binding pattern as it failed to stain cells infected with DENV-1, DENV-2, DENV-3 or DENV-4, or bind to purified WNV E protein (Extended Data Fig. 3 and data not shown). In comparison, ZIKV-116 bound to cells infected with DENV-1, DENV-2 or DENV-4, but did not bind to DENV-2 DIII or WNV DIII in ELISA.

In vivo models of ZIKV pathogenesis and antibody prophylaxis have been reported^{7,10,11} in mice deficient in type-I interferon signalling. To determine whether ZIKV-117 had therapeutic activity, we treated 4–5-week-old wild-type male C57BL/6 mice at day -1 with anti-*Ifnar1* mAbs, and then inoculated animals with 10^3 focus-forming units (FFU) of a mouse-adapted African strain of ZIKV-Dakar⁷. Animals were treated with a single dose of ZIKV-117 or non-binding isotype control (human (h)CHK-152)¹², on day +1 (100 μ g; 6.7 mg kg⁻¹) or day +5 (250 μ g; 16.7 mg kg⁻¹). Animals treated with hCHK-152 sustained significant lethality compared to those receiving ZIKV-117 (Fig. 3a), which were protected even when administered only a single dose 5 days after virus inoculation.

We and others have demonstrated placental injury and fetal demise following ZIKV infection of pregnant mice with deficiencies in type-I interferon signalling^{13–15}. To assess the protective ability of ZIKV-117 during fetal development, we treated *Ifnar1*^{-/-} pregnant dams mated to wild-type male mice with a single 250 μ g dose of ZIKV-117 or isotype control mAb (hCHK-152) on embryo day 5.5 (E5.5), the day before ZIKV inoculation. Whereas inoculation with ZIKV-Brazil at E6.5 following administration with hCHK-152 resulted in high levels of maternal infection and almost uniform fetal demise by E13.5, treatment with ZIKV-117 improved fetal outcome (Fig. 3b, c).

Because of the extent of demise at E13.5 after ZIKV infection of *Ifnar1*^{-/-} dams, we could not recover adequate numbers of fetuses to measure viral titres. Accordingly, we switched to a wild-type mouse model with an acquired type-I interferon deficiency using the mouse-adapted African ZIKV-Dakar strain. Wild-type pregnant dams were treated at day -1 (E5.5) with an anti-*Ifnar1* mAb. At the same time, these animals were administered vehicle control (PBS), 250 μ g isotype control hCHK-152, or 250 μ g ZIKV-117. One day later, dams were inoculated subcutaneously with 10^3 FFU of ZIKV-Dakar. Fetuses from dams treated with anti-*Ifnar1* mAbs and given PBS or hCHK-152 showed high levels (for example, around 10^5 to 10^7 FFU equivalents per gram) of viral RNA in the placenta and fetal brain (Fig. 3d). In comparison, mice treated with anti-*Ifnar1* and ZIKV-117 had reduced virus levels in the placenta and fetal brain (for example, around 10^0 to 10^3 FFU equivalents per gram). This phenotype was associated with transport of human ZIKV E-specific IgG across the maternal–fetal placental barrier (816 ± 53 ng ml⁻¹ for the placenta and $1,675 \pm 203$ ng ml⁻¹ for the fetal head; Extended Data Fig. 4). As levels of neonatal Fc receptor in the mouse placenta are lower than other mammalian species¹⁶, reduced levels of transport of maternal or exogenous IgG into the fetus is expected¹⁷. Although this factor could underestimate the therapeutic effect of exogenous anti-ZIKV IgG or maternal antibodies, we nonetheless achieved levels in the placenta and fetal head that were orders of magnitude above the IC₅₀ neutralization value for ZIKV-117.

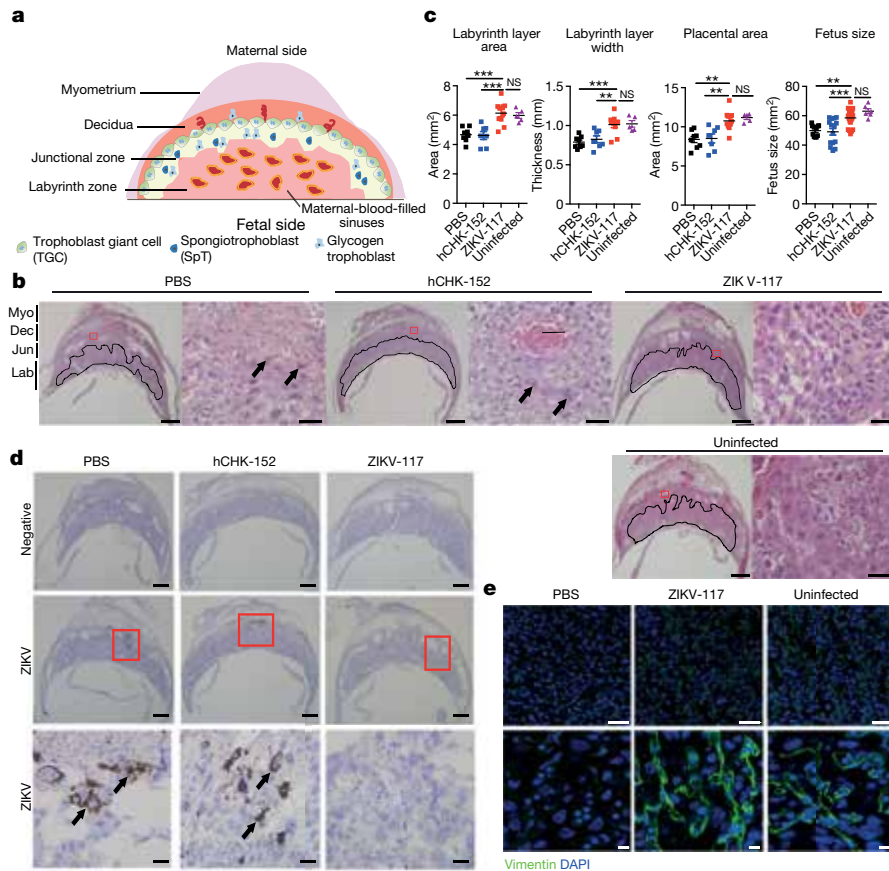


Figure 4 | Effect of ZIKV-117 treatment on the placenta and the fetus.

a, Cartoon depicting murine placental structures and zones. **b–e**, Pregnant dams were treated with PBS, hCHK-152, or ZIKV-117 as described in Fig. 3d–f before infection with ZIKV-Dakar or mock-infected. **b**, Haematoxylin and eosin staining of placenta at E13.5. Placental labyrinth zone is marked with a solid line. Low power (scale bar, 1 mm) and high power (scale bar, 50 μ m) images are presented in sequence. Black arrows indicate apoptotic trophoblasts in areas corresponding to regions of ZIKV infectivity (see panel **d**, below). **c**, Measurements of thickness and indicated areas of placenta and fetus body size. Each symbol represents data from an individual placenta or fetus. Significance was analysed by

ANOVA with a Dunn's multiple comparison test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, $P > 0.05$, NS, not significant). **d**, *In situ* hybridization. Low (scale bar, 500 μ m) and high (scale bar, 50 μ m) power images are presented in sequence. Black arrows indicate cells positive for ZIKV RNA in the junctional zone of the placenta. The images in panels are representative of several placentas from independent dams. **e**, Low (scale bar, 50 μ m) and high (scale bar, 10 μ m) power magnified images of immunofluorescence staining of placentas for vimentin (in green, which marks fetal capillary endothelium) from ZIKV-infected dams treated with PBS or ZIKV-117 or from uninfected pregnant animals. Nuclei are counter-stained blue with DAPI.

Dams treated with ZIKV-117 also had substantially lower levels of viral RNA in the maternal brain and serum (Fig. 3e).

Antibody-dependent enhancement of flavivirus infection occurs when type-specific or cross-reactive antibodies fail to reach a stoichiometric threshold for neutralization and instead facilitate infection of Fc γ R-expressing myeloid cells¹⁸. Because antibodies can promote antibody-dependent enhancement of ZIKV in cell culture^{19,20}, we evaluated the protective efficacy of a recombinant form of ZIKV-117 IgG containing a leucine (L) to alanine (A) substitution at positions 234 and 235 (LALA)²¹, which lacked efficient binding to Fc γ R, retained interactions with FcRn²², and neutralized ZIKV *in vitro* equivalently compared to the parent mAb (Extended Data Fig. 5). The LALA variant of ZIKV-117 showed similar protective activity against infection of the placenta and fetus relative to the parent mAb (Fig. 3f). As the protection conferred by ZIKV-117 in the pregnancy model is probably due to neutralization and not Fc effector functions, LALA variants could be used without a risk of antibody-dependent enhancement.

We next assessed the post-exposure efficacy of ZIKV-117 during pregnancy. Mice treated with anti-Ifnar1 mAbs at E5.5 were inoculated with 10³ FFU of ZIKV-Dakar at E6.5 and then administered a single dose of PBS, 250 μ g of hCHK-152, or 250 μ g of ZIKV-117 at E7.5. Compared to PBS or isotype control mAb treatment, administration of ZIKV-117 markedly reduced the viral burden in the dams, the placenta and fetus when measured at E13.5 (Fig. 3g, h).

The reduction in viral load mediated by ZIKV-117 was associated with decreased damage of the placenta (as judged by labyrinth layer and overall placenta area), less trophoblast cell death, and increased body size of the fetus (Fig. 4a–c) compared to fetuses of PBS- or hCHK-152-treated dams. ZIKV-117 protected against ZIKV-induced placental insufficiency, as the placental area and fetal size from infected dams treated with anti-ZIKV mAbs were similar to that of uninfected placentas¹⁴. *In situ* hybridization revealed an almost complete absence of viral RNA in the junctional zone and decidua of the placenta in animals treated with ZIKV-117 compared to staining observed in PBS- or hCHK-152-treated controls (Fig. 4d, Extended Data Fig. 6). We also observed vascular damage associated with ZIKV infection of the placenta¹⁴, characterized as diminished vimentin staining of fetal endothelial cells, which was rescued by ZIKV-117 to levels seen in uninfected placentas (Fig. 4e). The histopathological data suggests that ZIKV-117 treatment can reduce the ability of ZIKV to cross the fetal endothelial cell barrier, and thereby prevent vertical transmission and improve fetal outcome.

Our most potent neutralizing antibodies exhibited a breadth of inhibitory activity against strains from Africa, Asia, and the Americas. Even a single ZIKV-117 dose given 5 days after infection protected mice against lethal infection, a timeline similar to the most protective antibodies against other flaviviruses²³. Prophylaxis or post-exposure therapy of pregnant mice with ZIKV-117 reduced infection in mothers,

and in placental and fetal tissues. As the extent to which these observations in mice translate to humans remains unclear, protection studies in non-human primates, which share a placental architecture similar to humans, seem warranted. If the results were consistent, ZIKV-117 or human antibodies with similar profiles^{10,19} could be developed as a treatment measure during pregnancy for at-risk humans. By defining key epitopes on the E protein associated with antibody-mediated protection, our studies also inform vaccine efforts to design new immunogens that elicit highly protective antibody responses against ZIKV.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 October; accepted 27 October 2016.

Published online 7 November 2016.

- Coyne, C. B. & Lazear, H. M. Zika virus—reigniting the TORCH. *Nat. Rev. Microbiol.* **14**, 707–715 (2016).
- Oehler, E. *et al.* Zika virus infection complicated by Guillain-Barré syndrome—case report, French Polynesia, December 2013. *Eur. Commun. Dis. Bull.* **19**, 7–9 (2014).
- Musso, D., Nilles, E. J. & Cao-Lormeau, V. M. Rapid spread of emerging Zika virus in the Pacific area. *Clin. Microbiol. Infect.* **20**, 0595–0596 (2014).
- Araujo, A. Q. C., Silva, M. T. T. & Araujo, A. P. Q. C. Zika virus-associated neurological disorders: a review. *Brain* **139**, 2122–2130 (2016).
- Gatherer, D. & Kohl, A. Zika virus: a previously slow pandemic spreads rapidly through the Americas. *J. Gen. Virol.* **97**, 269–273 (2016).
- Davidson, E. & Doranz, B. J. A high-throughput shotgun mutagenesis approach to mapping B-cell antibody epitopes. *Immunology* **143**, 13–20 (2014).
- Zhao, H. *et al.* Structural basis of Zika virus-specific antibody protection. *Cell* **166**, 1016–1027 (2016).
- Dowd, K. A., DeMaso, C. R. & Pierson, T. C. Genotypic differences in dengue virus neutralization are explained by a single amino acid mutation that modulates virus breathing. *MBio* **6**, e01559–159 (2015).
- Dowd, K. A., Mukherjee, S., Kuhn, R. J. & Pierson, T. C. Combined effects of the structural heterogeneity and dynamics of flaviviruses on antibody recognition. *J. Virol.* **88**, 11726–11737 (2014).
- Stettler, K. *et al.* Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection. *Science* **353**, 823–826 (2016).
- Swanstrom, J. A. *et al.* Dengue virus envelope dimer epitope monoclonal antibodies isolated from dengue patients are protective against Zika virus. *MBio* **7**, e01123–16 (2016).
- Pal, P. *et al.* Development of a highly protective combination monoclonal antibody therapy against Chikungunya virus. *PLoS Pathog.* **9**, e1003312 (2013).
- Mysorekar, I. U. & Diamond, M. S. Modeling Zika virus infection in pregnancy. *N. Engl. J. Med.* **375**, 481–484 (2016).
- Miner, J. J. *et al.* Zika virus infection during pregnancy in mice causes placental damage and fetal demise. *Cell* **165**, 1081–1091 (2016).
- Yockey, L. J. *et al.* Vaginal exposure to Zika virus during pregnancy leads to fetal brain infection. *Cell* **166**, 1247–1256.e4 (2016).
- Kim, J. *et al.* FcRn in the yolk sac endoderm of mouse is required for IgG transport to fetus. *J. Immunol.* **182**, 2583–2589 (2009).
- Pentšuk, N. & van der Laan, J. W. An interspecies comparison of placental antibody transfer: new insights into developmental toxicity testing of monoclonal antibodies. *Birth Defects Res. B Dev. Reprod. Toxicol.* **86**, 328–344 (2009).
- Pierson, T. C. *et al.* The stoichiometry of antibody-mediated neutralization and enhancement of West Nile virus infection. *Cell Host Microbe* **1**, 135–145 (2007).
- Dejnirattisai, W. *et al.* Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with Zika virus. *Nat. Immunol.* **17**, 1102–1108 (2016).
- Charles, A. S. & Christofferson, R. C. Utility of a dengue-derived monoclonal antibody to enhance Zika infection *in vitro*. *PLoS Curr.* **8**, 1–31 (2016).
- Hessell, A. J. *et al.* Fc receptor but not complement binding is important in antibody protection against HIV. *Nature* **449**, 101–104 (2007).
- Oliphant, T. *et al.* Development of a humanized monoclonal antibody with therapeutic potential against West Nile virus. *Nat. Med.* **11**, 522–530 (2005).

Acknowledgements We thank N. Murphy, J. Govero, M. Gorman, J. Miner, R. Fong and S. Reddy for technical help and advice on experiments. This work was supported by US N.I.H. grants R01 AI073755 (to M.S.D., D.H.F. and J.E.C.), R01 AI104972 (to M.S.D.), US N.I.H. contracts HHSN272201400024C (to J.E.C.), HHSN272201400058C (to B.J.D.), HHSN272201400018C (to D.H.F., M.S.D. and J.E.C.) and HHSN272201200026C (CSGID; to D.H.F.), and by a Preventing Prematurity Initiative grant from the Burroughs Wellcome Fund and an Investigator award from the March of Dimes (to I.U.M.). E.F. was supported by an N.I.H. Pre-doctoral training grant award (T32 AI007163).

Author Contributions G.S., E.F., I.U.M., B.J.D., M.S.D. and J.E.C. planned the studies. G.S., E.F., N.K., J.M.F., R.G.B., B.C., A.L.B., T.B. and E.D. conducted experiments. H.Z., C.A.N. and D.H.F. provided protein reagents. G.S., E.F., M.S.D., B.C., B.J.D., I.U.M. and J.E.C. interpreted the studies. G.S., E.F., M.S.D. and J.E.C. wrote the first draft of the paper. D.H.F., B.J.D., M.S.D. and J.E.C. obtained funding. All authors reviewed, edited and approved the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.E.C. (james.crowe@vanderbilt.edu) or M.S.D. (diamond@wustl.wustl.edu).

METHODS

Research subjects. We studied eight subjects in the United States with previous or recent ZIKV infection (Extended Data Table 2). The studies were approved by the Institutional Review Board of Vanderbilt University Medical Center; samples were obtained after informed consent was obtained by the Vanderbilt Clinical Trials Center. Two subjects (972 and 973) were infected with an African lineage strain in 2008 (one subject while working in Senegal, the second acquired the infection by sexual transmission from the first, as previously reported²⁴). The other six subjects were infected during the current outbreak of an Asian lineage strain, following exposure in Brazil, Mexico or Haiti.

Generation and quantification of human B-cell lines secreting ZIKV E protein specific antibodies. Peripheral blood mononuclear cells (PBMCs) from heparinized blood were isolated with Ficoll-Histopaque by density gradient centrifugation. The cells were used immediately or cryopreserved in the vapour phase of liquid nitrogen until use. Ten million PBMCs were cultured in 384-well plates (Nunc) using culture medium (ClonaCell-HY Medium A, StemCell Technologies) supplemented with $8\mu\text{g ml}^{-1}$ of the TLR agonist CpG (phosphorothioate-modified oligodeoxynucleotide ZOEZOEZZZZOEEZOEZZZT, Invitrogen), $3\mu\text{g ml}^{-1}$ of Chk2 inhibitor (Sigma), $1\mu\text{g ml}^{-1}$ of cyclosporine A (Sigma), and clarified supernatants from cultures of B95.8 cells (ATCC) containing Epstein–Barr virus. After 7 days, cells from each 384-well culture plate were expanded into four 96-well culture plates (Falcon) using ClonaCell-HY Medium A containing $8\mu\text{g ml}^{-1}$ of CpG, $3\mu\text{g ml}^{-1}$ of Chk2 inhibitor, and 10^7 irradiated heterologous human PBMCs (Nashville Red Cross) and cultured for an additional 4 days. Supernatants were screened in ELISA (described below) for reactivity with various ZIKV E proteins, which are described below. The minimal frequency of ZIKV E-reactive B cells was estimated based on the number of wells with E protein-reactive supernatants compared with the total number of lymphoblastoid cell line colonies in the transformation plates (calculation: E-reactive B-cell frequency = (number of wells with E-reactive supernatants) divided by (number of LCL colonies in the plate) $\times 100$).

Protein expression and purification. The ectodomains of ZIKV E (H/PP/2013; GenBank Accession KJ776791) and the fusion-loop mutant E-FLM (containing four mutations: T76A, Q77G, W101R, L107R) were expressed transiently in Expi293F cells and purified as described previously⁷. ZIKV DIII (residues 299–407 of strain H/PP/2013), WNV DIII (residues 296–405 of strain New York 1999) and DENV-2 DIII (residues 299–410 of strain 16681) were expressed in BL21 (DE3) as inclusion bodies and refolded *in vitro*²⁵. Briefly, inclusion bodies were denatured and refolded by gradual dilution into a refolding buffer (400 mM L-arginine, 100 mM Tris (pH 8.3), 2 mM EDTA, 5 and 0.5 mM reduced and oxidized glutathione) at 4°C. Refolded proteins were purified by size-exclusion chromatography using a Superdex 75, 16/60 (GE Healthcare).

Generation of human hybridomas. Cells from wells with transformed B cells containing supernatants that exhibited reactivity to ZIKV E protein were fused with HMM2.5 myeloma cells (gift from L. Cavacini) using an established electrofusion technique²⁶. After fusion, hybridomas were suspended in a selection medium containing 100 μM hypoxanthine, 0.4 μM aminopterin, 16 μM thymidine (HAT Media Supplement, Sigma), and 7 $\mu\text{g ml}^{-1}$ ouabain (Sigma) and cultured in 384-well plates for 18 days before screening hybridomas for antibody production by ELISA. After fusion with HMM2.5 myeloma cells, hybridomas producing ZIKV E-specific antibodies were cloned biologically by single-cell fluorescence-activated cell sorting. Hybridomas were expanded in post-fusion medium (ClonaCell-HY Medium E, STEMCELL Technologies) until 50% confluent in 75-cm² flasks (Corning).

For antibody production, cells from one 75-cm² flask were collected with a cell scraper and expanded to four 225-cm² flasks (Corning) in serum-free medium (Hybridoma-SFM, Life Technologies). After 21 days, supernatants were clarified by centrifugation and filtered using 0.45- μm pore size filter devices. HiTrap Protein G or HiTrap MabSelectSure columns (GE Healthcare Life Sciences) were used to purify antibodies from filtered supernatants.

Sequence analysis of antibody variable region genes. Total cellular RNA was extracted from pelleted cells from hybridoma clones, and an RT-PCR reaction was performed using mixtures of primers designed to amplify all heavy-chain or light-chain antibody variable regions²⁷. The generated PCR products were purified using AMPure XP magnetic beads (Beckman Coulter) and sequenced directly using an ABI3700 automated DNA sequencer. The variable region sequences of the heavy and light chains were analysed using the IMGT/V-Quest program^{28,29}.

ELISA and EC₅₀ binding analysis. Wells of microtitre plates were coated with purified, recombinant ectodomain of ZIKV E, DIII, DIII-LR mutants (DIII containing A310E and T335K mutations) or DIII of related flaviviruses DENV-2 or WNV and incubated at 4°C overnight. In ELISA studies with purified mAbs, we used recombinant ZIKV E protein ectodomain with His₆ tag produced in Sf9 insect cells (Meridian Life Sciences R01635). Plates were blocked with 5% skimmed

milk in PBS-T for 1 h. B-cell culture supernatants or purified antibodies were added to the wells and incubated for 1 h at ambient temperature. The bound antibodies were detected using goat anti-human IgG (γ -specific) conjugated with alkaline phosphatase (Southern Biotech) and pNPP disodium salt hexahydrate substrate (Sigma). In ELISAs that assessed binding of mAbs to DIII and DIII LR mutants, we used previously described murine mAbs ZV-2 and ZV-54 (ref. 7) as controls. A goat anti-mouse IgG conjugated with alkaline phosphatase (Southern Biotech) was used for detection of these antibodies. Colour development was monitored at 405 nm in a spectrophotometer (Biotek). For determining EC₅₀, microtitre plates were coated with ZIKV E or E-FLM that eliminated interaction of fusion-loop specific antibodies. Purified antibodies were diluted serially and applied to the plates. Bound antibodies were detected as above. A nonlinear regression analysis was performed on the resulting curves using Prism (GraphPad) to calculate EC₅₀ values.

ELISA for detection of human antibodies in murine tissues. Fetal head and placental tissues were collected at E13.5 from groups treated with ZIKV-117 or PBS (as a negative control), homogenized in PBS (250 μl) and stored at -20°C . ELISA plates were coated with ZIKV E protein, and thawed, clarified tissue homogenates were applied undiluted in triplicate. Bound antibodies were detected using goat anti-human IgG (Fc-specific) antibody conjugated with alkaline phosphatase. The quantity of antibody was determined by comparison with a standard curve constructed using purified ZIKV-117 in a dilution series.

Biolayer interferometry competition binding assay. His₆-tagged ZIKV E protein was immobilized on anti-His coated biosensor tips (Pall) for 2 min on an Octet Red biosensor instrument. After measuring the baseline signal in kinetics buffer (PBS, 0.01% BSA, and 0.002% Tween 20) for 1 min, biosensor tips were immersed into the wells containing first antibody at a concentration of $10\mu\text{g ml}^{-1}$ for 7 min. Biosensors then were immersed into wells containing a second mAb at a concentration of $10\mu\text{g ml}^{-1}$ for 7 min. The signal obtained for binding of the second antibody in the presence of the first antibody was expressed as a percentage of the uncompleted binding of the second antibody that was derived independently. The antibodies were considered competing if the presence of first antibody reduced the signal of the second antibody to less than 30% of its maximal binding and non-competing if the signal was greater than 70%. A level of 30–70% was considered intermediate competition.

Shotgun mutagenesis epitope mapping. Epitope mapping was performed by shotgun mutagenesis essentially as described previously⁶. A ZIKV prM/E protein expression construct (based on ZIKV strain SPH2015) was subjected to high-throughput alanine scanning mutagenesis to generate a comprehensive mutation library. Each residue within prM/E was changed to alanine, with alanine codons mutated to serine. In total, 672 ZIKV prM/E mutants were generated (100% coverage), sequence confirmed, and arrayed into 384-well plates. Each ZIKV prM/E mutant was transfected into HEK-293T cells and allowed to express for 22 h. Cells were fixed in 4% (v/v) paraformaldehyde (Electron Microscopy Sciences), and permeabilized with 0.1% (w/v) saponin (Sigma-Aldrich) in PBS plus calcium and magnesium (PBS++). Cells were incubated with purified mAbs diluted in PBS++, 10% normal goat serum (Sigma), and 0.1% saponin. Primary antibody screening concentrations were determined using an independent immunofluorescence titration curve against wild-type ZIKV prM/E to ensure that signals were within the linear range of detection. Antibodies were detected using 3.75 $\mu\text{g ml}^{-1}$ of AlexaFluor488-conjugated secondary antibody (Jackson ImmunoResearch Laboratories) in 10% NGS/0.1% saponin. Cells were washed three times with PBS++/0.1% saponin followed by two washes in PBS. Mean cellular fluorescence was detected using a high-throughput flow cytometer (HTFC, Intellicyt). Antibody reactivity against each mutant prM/E clone was calculated relative to wild-type prM/E protein reactivity by subtracting the signal from mock-transfected controls and normalizing to the signal from wild-type prM/E-transfected controls. Mutations within clones were identified as critical to the mAb epitope if they did not support reactivity of the test mAb, but supported reactivity of other ZIKV antibodies. This counter-screen strategy facilitates the exclusion of prM/E mutants that are locally misfolded or have an expression defect.

Vertebrate animal studies ethics statement. This study was carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The protocols were approved by the Institutional Animal Care and Use Committee at the Washington University School of Medicine (Assurance number A3381-01). Inoculations were performed under anaesthesia induced and maintained with ketamine hydrochloride and xylazine, and all efforts were made to minimize animal suffering. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Viruses and cells. ZIKV strain H/PP/2013 (French Polynesia, 2013) was obtained from X. de Lamballerie (Aix Marseille Université). ZIKV Brazil Paraiba 2015

was provided by S. Whitehead (Bethesda) and originally obtained from P. F. C. Vasconcelos (Instituto Evandro Cargas). ZIKV MR 766 (Uganda, 1947), Malaysia P6740 (1966), and Dakar 41519 (Senegal, 1982) were provided by the World Reference Center for Emerging Viruses and Arboviruses (R. Tesh, University of Texas Medical Branch). Nicaraguan DENV strains (DENV-1 1254-4, DENV-2 172-08, DENV-3 N2845-09, and DENV-4 N703-99) were provided generously by E. Harris (University of California, Berkeley). Virus stocks were propagated in C6/36 *Aedes albopictus* cells (DENV) or Vero cells (ZIKV). ZIKV Dakar 41519 (ZIKV-Dakar) was passaged twice *in vivo* in *Rag1*^{-/-} mice (M. Gorman and M. Diamond, unpublished data) to create a mouse-adapted strain. Virus stocks were titrated by focus-forming assay (FFA) on Vero cells. All cell lines were checked regularly for mycoplasma contamination and were negative. Cell lines were authenticated at acquisition with short tandem repeat method profiling; Vero cells, though commonly misidentified in the field, were used as they are the standard cell line for flavivirus titration.

Neutralization assays. Serial dilutions of mAbs were incubated with 10² FFU of different ZIKV strains (MR 766, Dakar 41519, Malaysia P6740, H/PF/2013, or Brazil Paraiba 2015) for 1 h at 37 °C. The mAb-virus complexes were added to Vero cell monolayers in 96-well plates for 90 min at 37 °C. Subsequently, cells were overlaid with 1% (w/v) methylcellulose in MEM supplemented with 4% heat-inactivated FBS. Plates were fixed 40 h later with 1% PFA in PBS for 1 h at room temperature. The plates were incubated sequentially with 500 ng ml⁻¹ mouse anti-ZIKV (ZV-16, E.F. and M.S.D., unpublished data) and horseradish-peroxidase-conjugated goat anti-mouse IgG in PBS supplemented with 0.1% (w/v) saponin (Sigma) and 0.1% BSA. ZIKV-infected cell foci were visualized using TrueBlue peroxidase substrate (KPL) and quantitated on an ImmunoSpot 5.0.37 macroanalyzer (Cellular Technologies).

mAb binding to ZIKV- or DENV-infected cells. C6/36 *Aedes albopictus* cells were inoculated with a MOI 0.01 of ZIKV (H/PF/2013) or different DENV serotypes (Nicaraguan strains DENV-1 1254-4, DENV-2 172-08, DENV-3 N2845-09, DENV-4 N703-99). At 120 h post infection, cells were fixed with 4% PFA diluted in PBS for 20 min at room temperature and permeabilized with HBSS supplemented with 10 mM HEPES, 0.1% saponin and 0.025% NaN₃ for 10 min at room temperature. 50,000 cells were transferred to U-bottom plates and incubated for 30 min at 4 °C with 5 µg ml⁻¹ of anti-ZIKV human mAbs or negative (hCHK-152)¹², or positive (hE60)³⁰ isotype controls. After washing, cells were incubated with Alexa-Fluor-647-conjugated goat anti-human IgG (Invitrogen) at 1:500, fixed in 1% PFA in PBS, processed on MACSQuant Analyzed (Miltenyi Biotec), and analysed using FlowJo software (Tree Star).

Recombinant antibody expression and purification. Total RNA was extracted from hybridoma cells and genes encoding the VH and VL domains were amplified in RT-PCR using IgExp primers³¹. The PCR products were directly cloned into antibody expression vectors containing the constant domains of wild-type γ 1 chain, LALA mutant (leucine (L) to alanine (A) substitution at positions 234 and 235) γ 1 chain for the VH domains, and wild-type κ chain for the VL domain in an isothermal amplification reaction (Gibson reaction)³². Plasmids encoding the heavy and light chain were transfected into 293F cells and full-length recombinant IgG was secreted into transfected cell supernatants. Supernatants were collected and IgG purified using Protein G chromatography and eluted into PBS. The functional abrogation of the binding of the LALA variant IgG was confirmed in an ELISA binding assay with recombinant human Fc γ RI. The binding of wild-type ZIKV-117 or LALA antibody to Fc γ RI was evaluated, in comparison with the binding pattern of control antibodies (human mAb CKV063 (ref. 33) LALA mutated IgG).

Adult mouse lethal protection experiments. C57BL/6 male mice (4–5-week-old, Jackson Laboratories) were inoculated with 10³ FFU of mouse-adapted ZIKV-Dakar by subcutaneous route in the footpad. One-day before infection, mice were treated with 2 mg anti-Ifnar1 mAb (MAR1-5A3, Leinco Technologies) by intraperitoneal injection. ZIKV-specific human mAb (ZIKV-117) or an isotype control (hCHK-152) was administered as a single dose at day +1 (100 µg) or day +5 (250 µg) after infection through an intraperitoneal route. Animals were monitored for 21 days.

Pregnant mouse protection experiments. Wild-type C57BL/6 mice were bred in a specific pathogen-free facility at Washington University School of Medicine.

(1) *Ifnar1*^{-/-} dams, prophylaxis studies: *Ifnar1*^{-/-} female and wild-type male mice were mated; at E5.5, dams were treated with a single 250 µg dose of ZIKV mAb or isotype control by intraperitoneal injection. At E6.5, mice were inoculated with 10³ FFU of ZIKV Brazil Paraiba 2015 by subcutaneous injection in the footpad. (2) Wild-type dams, prophylaxis studies: wild-type female and male mice were mated; at embryonic days E5.5, dams were treated with a single 250 µg dose of ZIKV mAb or isotype control by intraperitoneal injection as well as a 1 mg injection of anti-Ifnar1 (MAR1-5A3). At E6.5, mice were inoculated with 10³ FFU of mouse-adapted ZIKV-Dakar by subcutaneous injection in the footpad. At E7.5, dams received a second 1 mg dose of anti-Ifnar1 through an intraperitoneal route.

(3) Wild-type dams, therapy studies: wild-type female and male mice were mated; at embryonic days E5.5, dams were treated with a 1 mg injection of anti-Ifnar1 (MAR1-5A3). At E6.5, mice were inoculated with mouse-adapted 10³ FFU of ZIKV-Dakar by subcutaneous injection in the footpad. At E7.5, dams received a second 1 mg dose of anti-Ifnar1 as well as a single 250 µg dose of ZIKV mAb or isotype control through an intraperitoneal route. All animals were euthanized at E13.5, and placentas, fetuses and maternal tissues were collected. Fetus size was measured as the crown-rump length \times occipitofrontal diameter of the head.

Measurement of viral burden. ZIKV-infected tissues were weighed and homogenized with stainless steel beads in a Bullet Blender instrument (Next Advance) in 200 µl of PBS. Samples were clarified by centrifugation (2,000g for 10 min). All homogenized tissues from infected animals were stored at -20 °C. Tissue samples and serum from ZIKV-infected mice were extracted with RNeasy 96 Kit (tissues) or Viral RNA Mini Kit (serum) (Qiagen). ZIKV RNA levels were determined by TaqMan one-step quantitative reverse transcriptase PCR (qRT-PCR) on an ABI7500 Fast Instrument using published primers and conditions³⁴. Viral burden was expressed on a log₁₀ scale as viral RNA equivalents per g or ml after comparison with a standard curve produced using serial tenfold dilutions of ZIKV RNA.

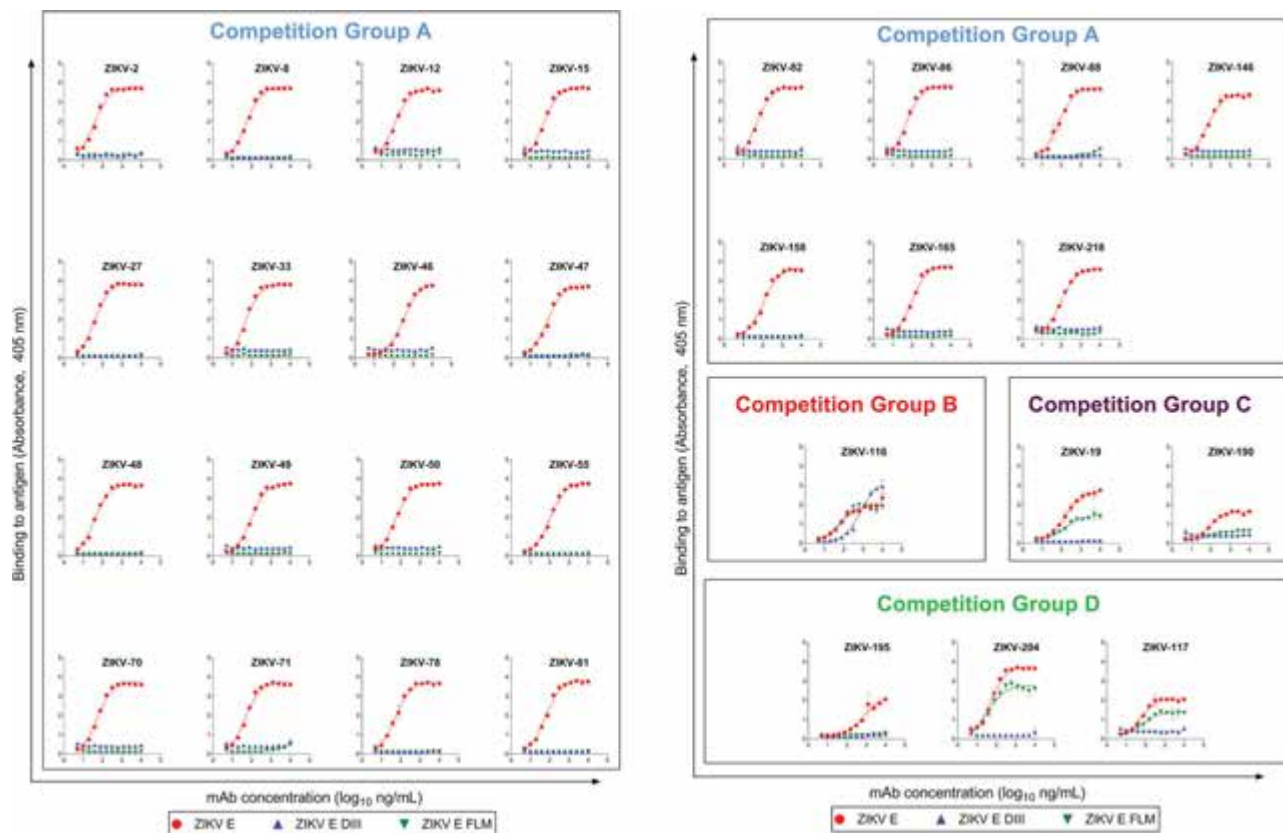
Viral RNA *in situ* hybridization. RNA *in situ* hybridization was performed with RNAscope 2.5 (Advanced Cell Diagnostics) according to the manufacturer's instructions. PFA-fixed paraffin embedded placental sections were deparaffinized by incubation for 60 min at 60 °C. Endogenous peroxidases were quenched with H₂O₂ for 10 min at room temperature. Slides were boiled for 15 min in RNAscope Target Retrieval Reagents and incubated for 30 min in RNAscope Protease Plus before probe hybridization. The probe targeting ZIKV RNA was designed and synthesized by Advanced Cell Diagnostics (catalogue number 467771). Negative (targeting bacterial gene *dapB*) control probes were also obtained from Advanced Cell Diagnostics (catalogue number 310043). Tissues were counterstained with Gill's haematoxylin and visualized with standard bright-field microscopy.

Histology and immunohistochemistry. Collected placentas were fixed in 10% neutral buffered formalin at room temperature and embedded in paraffin. At least three placentas from different litters with the indicated treatments were sectioned and stained with haematoxylin and eosin to assess morphology. Surface area and thickness of placenta and different layers were measured using Image J software. For immunofluorescence staining on mouse placentas, deparaffinized tissues were blocked in blocking buffer (1% BSA, 0.3% Triton, PBS) for 2 h and incubated with anti-vimentin antibody (1:500, rabbit, Abcam ab92547). Secondary antibody conjugated with Alexa 488 (1:500 in PBS) was applied for 1 h at room temperature. Samples were counterstained with DAPI (4',6'-diamidino-2-phenylindole, 1:1,000 dilution).

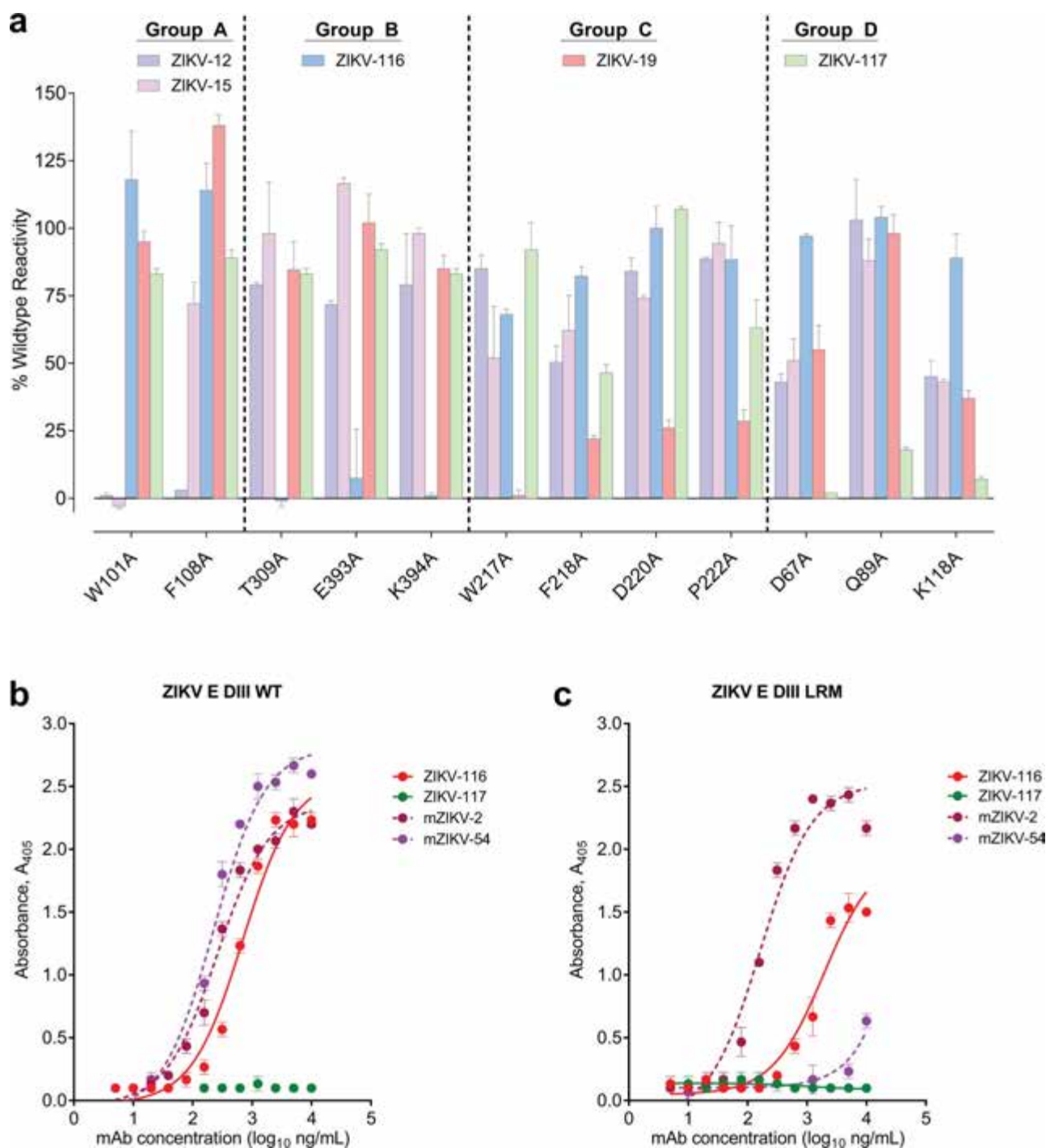
Statistical analysis. All virological data were analysed with GraphPad Prism software. Kaplan–Meier survival curves were analysed by the log rank test, and viraemia was compared using an ANOVA with a multiple comparisons test. *P* < 0.05 indicated statistically significant differences.

Data availability. All relevant data are included with the manuscript; source data for each of the main text figures is provided.

24. Foy, B. D. *et al.* Probable non-vector-borne transmission of Zika virus, Colorado, USA. *Emerg. Infect. Dis.* **17**, 880–882 (2011).
25. Nelson, C. A., Lee, C. A. & Fremont, D. H. Oxidative refolding from inclusion bodies. *Methods Mol. Biol.* **1140**, 145–157 (2014).
26. Yu, X., McGraw, P. A., House, F. S. & Crowe, J. E., Jr. An optimized electrofusion-based protocol for generating virus-specific human monoclonal antibodies. *J. Immunol. Methods* **336**, 142–151 (2008).
27. Thornburg, N. J. *et al.* Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J. Clin. Invest.* **123**, 4405–4409 (2013).
28. Brochet, X., Lefranc, M.-P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* **36**, W503–W508 (2008).
29. Giudicelli, V. & Lefranc, M. P. IMGT/junctionanalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (Ig) and T cell receptors (TR). *Cold Spring Harb. Protoc.* **2011**, 716–725 (2011).
30. Williams, K. L. *et al.* Therapeutic efficacy of antibodies lacking Fc γ R against lethal dengue virus infection is due to neutralizing potency and blocking of enhancing antibodies. *PLoS Pathog.* **9**, e1003157 (2013).
31. Thornburg, N. J. *et al.* H7N9 influenza virus neutralizing antibodies that possess few somatic mutations. *J. Clin. Invest.* **126**, 1482–1494 (2016).
32. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
33. Fong, R. H. *et al.* Exposure of epitope residues on the outer face of the chikungunya virus envelope trimer determines antibody neutralizing efficacy. *J. Virol.* **88**, 14364–14379 (2014).
34. Lanciotti, R. S. *et al.* Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. *Emerg. Infect. Dis.* **14**, 1232–1239 (2008).

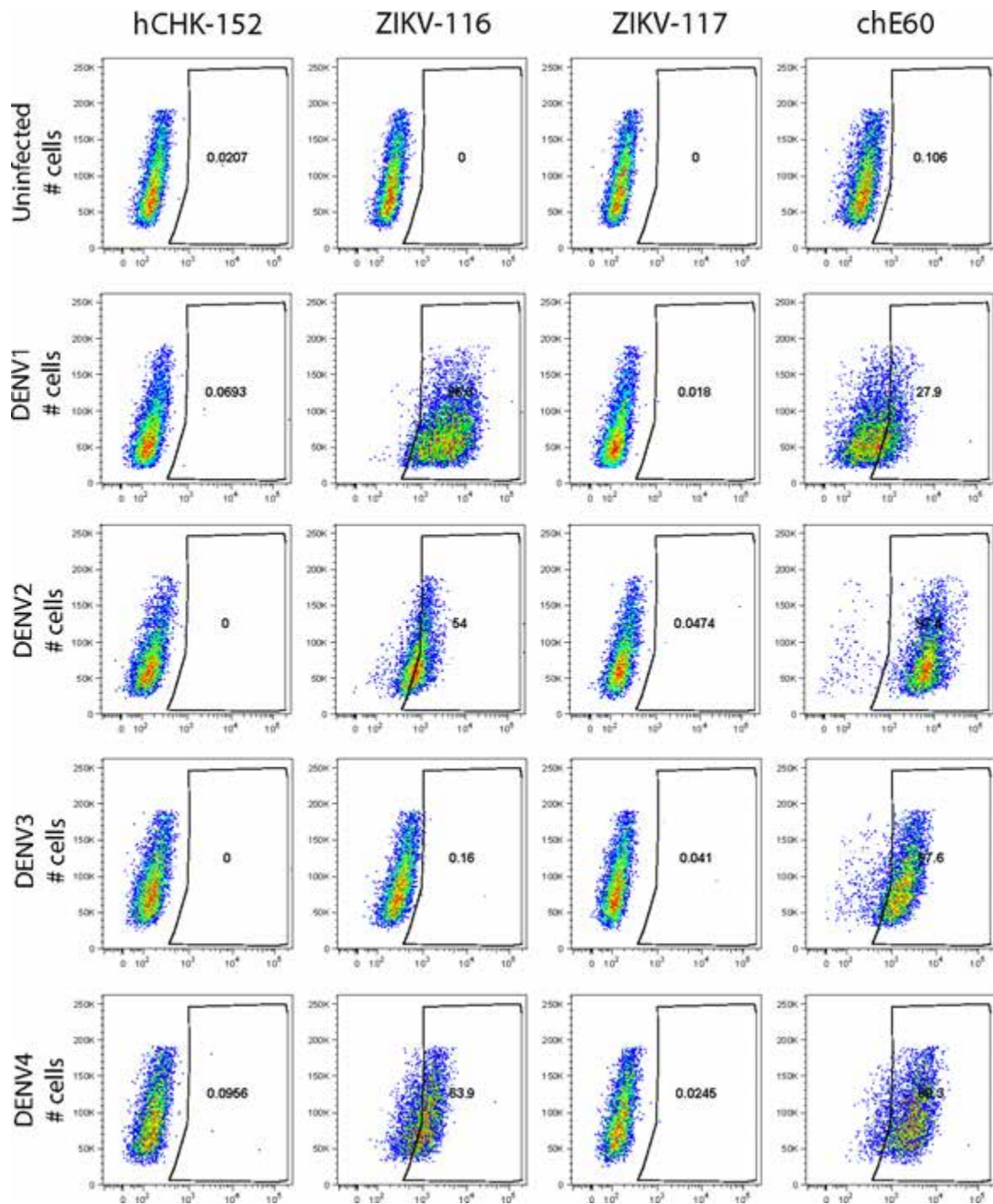


Extended Data Figure 1 | Binding of human mAbs to Zika E protein, E DIII or E-FLM. mAbs are organized by competition binding groups A to D. Purified mAbs were tested for binding to different antigens as indicated in ELISA as described in Methods. Non-linear regression analysis of the data was performed, and the data plotted are the mean and s.d.



Extended Data Figure 2 | High resolution epitope mapping of ZIKV mAbs. **a**, An alanine scanning mutation library for ZIKV envelope protein was constructed, in which each amino acid of prM/E was mutated individually to alanine (and alanine to serine) and expression constructs arrayed into 384-well plates, one mutation per well. Each clone in the ZIKV prM/E mutation library, expressed in HEK-293T cells, was tested for immunoreactivity with five mAbs from competition groups A–D, measured using an Intellicyt high-throughput flow cytometer. Shown here for each of the five mAbs is the reactivity with the ZIKV E protein mutants

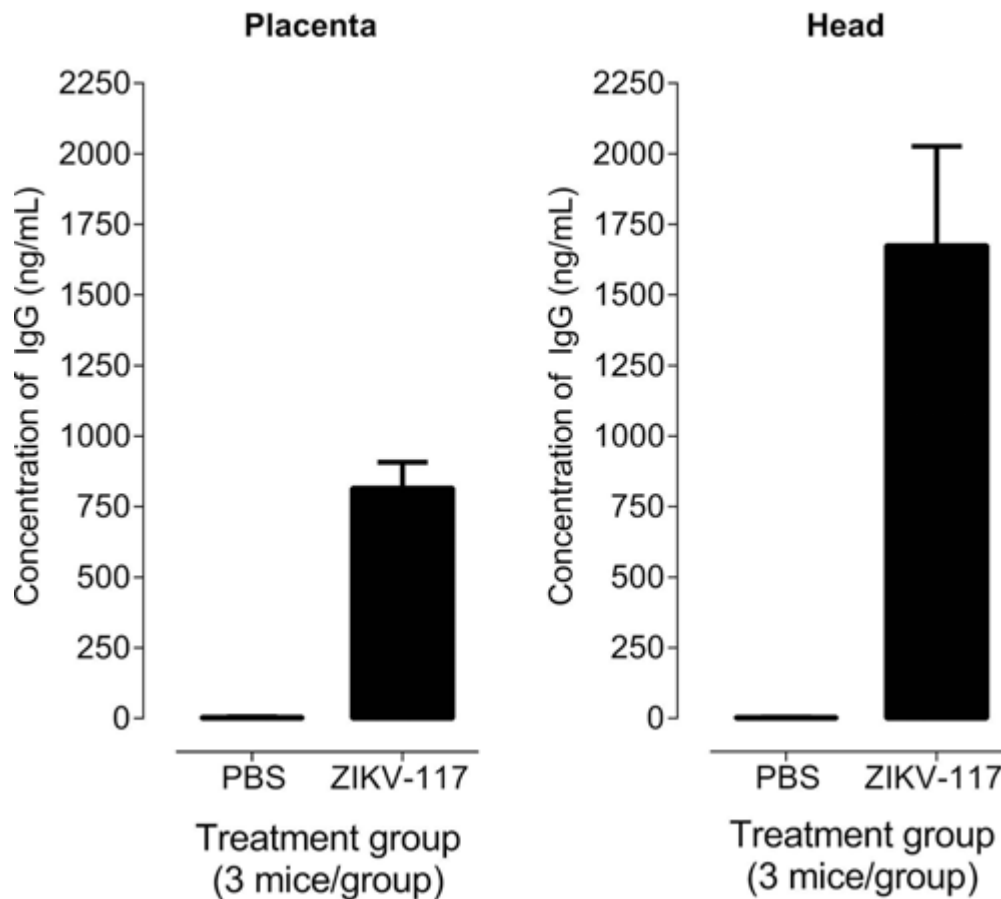
that identified the epitope residues for these mAbs. mAb reactivity for each alanine mutant are expressed as percent of the reactivity of mAb with wild-type ZIKV prM/E. Clones with reactivity <30% relative to wild-type ZIKV prM/E were identified as critical for mAb binding. Bars represent the mean and range of at least two replicate data points. Binding of group B mAbs, ZIKV-116 to wild-type ZIKV E DIII (**b**) or DIII LR mutant (**c**) was compared with mouse mAbs ZV-2 and ZV-54. Binding of ZIKV-116 was decreased by mutations in DIII-LR. Data plotted are mean \pm s.d.



Extended Data Figure 3 | Binding of human mAbs to permeabilized DENV-infected C6/36 cells. C6/36 cells were infected with DENV-1, DENV-2, DENV-3, DENV-4 or mock infected. Cells were stained with the indicated anti-ZIKV mAbs, an isotype negative control (hCHK-152),

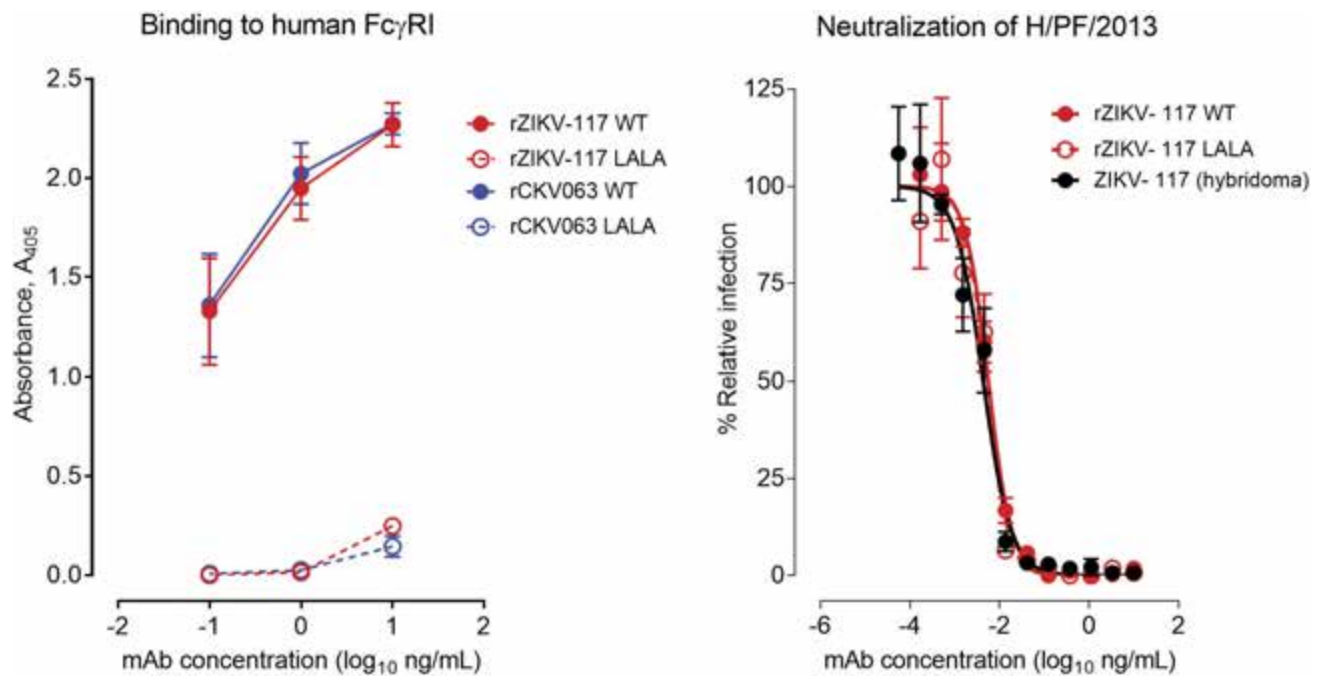
or a positive control (a cross-reactive antibody to DENV; chimeric human E60 (chE60)) and processed by flow cytometry. The data are representative of two independent experiments. The numbers in the box indicate the fraction of cells that stained positively.

Concentration of human IgG in fetal tissues



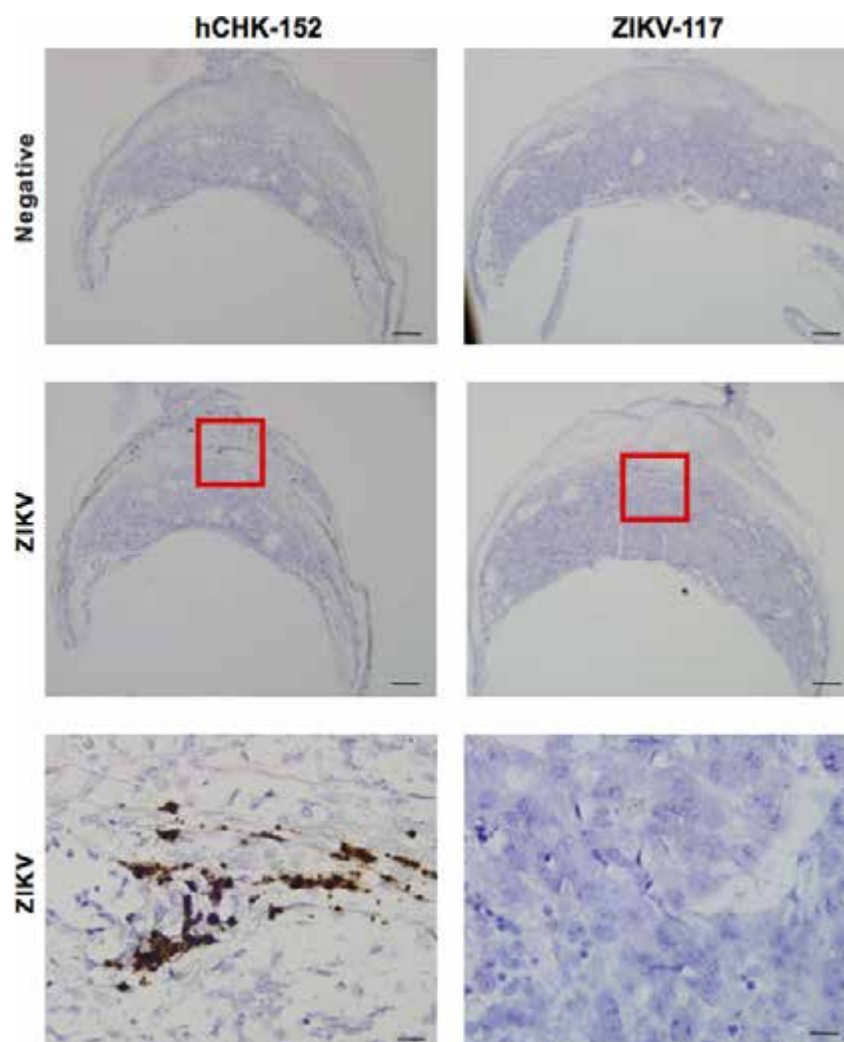
Extended Data Figure 4 | Detection of human IgG in placenta or fetal head tissues in ZIKV-117- or PBS-treated pregnant mice. As described in Fig. 3, wild-type female mice were mated with wild-type sires and monitored for pregnancy. At E5.5, dams were treated with anti-Ifnar1 mAb and PBS or 250 μ g of ZIKV-117. One day later (E6.5), dams were inoculated with 10^3 FFU of ZIKV-Dakar. Fetuses and placentas ($n = 4$ each) were collected on E13.5, homogenized, and tested for

human IgG by ELISA. Human antibody in tissues was captured on ELISA plates coated with ZIKV E protein and detected using goat anti-human IgG (Fc-specific) antibody. The quantity of antibody was determined by comparison with a standard curve constructed using purified ZIKV-117 in a dilution series. Four replicate measurements were performed for each mouse tissue and the results were averaged. The graphs represent the mean \pm s.e.m. from 3 mice per group.



Extended Data Figure 5 | Comparison of wild-type and LALA-mutated antibodies. **a**, Binding to recombinant human Fc γ RI. The functional abrogation of the binding of the LALA variant IgG was confirmed in an ELISA binding assay with recombinant human Fc γ RI. ZIKV-117 wild-type bound to Fc γ RI, whereas the ZIKV-117 LALA antibody did not. Wild-type and LALA versions of another human mAb, CKV063,

were used as controls. Binding to human Fc γ RI is one representative experiment of two, and error bars indicate s.e.m. of triplicate technical replicates. **b**, Neutralization assays. Wild-type ZIKV-117 and LALA antibodies exhibited equivalent neutralizing activity *in vitro* to each other and to the hybridoma-derived antibody. Neutralization assays are representative of two independent experiments completed in triplicate.



Extended Data Figure 6 | *In situ* hybridization of *Ifnar1*^{+/-} placenta after inoculation with ZIKV-Brazil and treatment with ZIKV-117. As described in Fig. 3a, *Ifnar1*^{-/-} female mice were mated with wild-type sires and monitored for pregnancy. At E5.5, dams were treated with 250 µg of either hCHK-152 isotype control or ZIKV-117. At E6.5, dams were inoculated with 10³ FFU of ZIKV-Brazil. Collected placentas were

fixed in 10% neutral buffered formalin at ambient temperature and embedded in paraffin. At least three placentas from different litters with the indicated treatments were sectioned for *in situ* hybridization staining using negative or ZIKV-specific RNA probes. Low (scale bar, 500 µm) and high (scale bar, 50 µm) power images are presented in sequence.

Extended Data Table 1 | Sequence characteristics of human mAbs

Clone	Isotype	Heavy chain				Light chain		
		V gene	J gene	D gene	HCDR lengths	V gene	J gene	LCDR lengths
ZIKV-2	IgG1, λ	V3-30-3*01	J3*02	D3-22*01	8.8.17	L2-11*01	J2*01	9.3.11
ZIKV-8	IgG1, κ	V3-30*04	J4*02	D3-22*01	8.8.15	K3-20*01	J1*01	7.3.9
ZIKV-12	IgG1, λ	V3-64*01	J4*02	D1-1*01	8.8.11	L7-46*01	J2*01	9.3.9
ZIKV-15	IgG1, κ	V3-30*02	J1*01	D3-22*01	8.8.21	K1-27*01	J4*01	6.3.9
ZIKV-19	IgG1, κ	V1-69*01	J4*03	D3-10*01	8.8.15	K1-27*01	J3*01	6.3.9
ZIKV-27	IgG1, κ	V1-18*01	J5*02	D6-19*01	8.8.13	K3-20*01	J3*01	7.3.11
ZIKV-33	IgG1, λ	V3-30*09	J2*01	D3-22*01	8.8.17	L1-51*01	J2*01	8.3.11
ZIKV-46	IgG1, κ	V3-15*01	J4*02	D3-22*01	8.10.14	K3-11*01	J1*01	6.3.8
ZIKV-47	IgG1, λ	V3-30-3*01	J1*01	D2-21*01	8.8.17	V1-40*01	J2*01	9.3.12
ZIKV-48	IgG1, λ	V3-30*14	J3*02	D4-17*01	8.8.12	L2-11*01	J2*01	9.3.9
ZIKV-49	IgG1, κ	V4-39*07	J3*02	D4-11*01	10.7.18	K1-12*01	J1*01	6.3.9
ZIKV-50	IgG1, κ	V3-72*01	J3*01	D6-6*01	8.10.11	K1-33*01	J4*01	6.3.8
ZIKV-55	IgG1, κ	V1-18*01	J4*02	D5-24*01	8.8.14	K1-9*01	J5*01	6.3.8
ZIKV-70	IgG1, λ	V3-30-3*01	J2*01	D3-22*01	8.8.17	L1-51*01	J2*01	8.3.10
ZIKV-71	IgG1, κ	V1-18*01	J6*02	D1-26*01	8.8.18	K1-9*01 F	J1*01	6.3.9
ZIKV-78	IgG1, λ	V3-23*04	J3*02	D3-22*01	8.8.20	L7-46*01	J3*02	9.3.10
ZIKV-81	IgG1, λ	V3-30-3*01	J1*01	D3-3*01	8.8.21	V1-40*01	J2*01	9.3.10
ZIKV-82	IgG1, κ	V4-39*07	J4*02	D3-16*01	10.7.21	K1-12*01	J1*01	6.3.9
ZIKV-86	IgG1, κ	V3-30*03	J5*02	D4-23*01	8.8.10	K1-5*03	J1*01	6.3.8
ZIKV-88	IgG1, λ	V3-30-3*01	J2*01	D3-22*01	8.8.17	L1-51*01	J2*01	8.3.11
ZIKV-116	IgG1, κ	V3-23*04	J4*02	D3-10*01	8.8.15	K1-5*03	J1*01	6.3.9
ZIKV-117	IgG1, κ	V3-30*02	J4*02	D3-10*01	8.8.12	K3-15*01	J1*01	6.3.9
ZIKV-146	IgG1, λ	V3-64*01	J4*02	D1-7*01	8.8.11	L7-46*01	J2*01	9.3.9
ZIKV-158	IgG1, λ	V3-30*04	J3*01	D4-17*01	8.8.14	V2-11*01	J2*01	9.3.11
ZIKV-165	IgG1, κ	V3-30*02	J1*01	D2-8*01	8.8.18	K1-5*03	J1*01	6.3.8
ZIKV-190	nd, λ	V3-23*04	J4*02	D3-3*01	8.8.13	L2-11*01	J2*01	9.3.11
ZIKV-195	nd, λ	V3-30*03	J6*02	D5-24*01	8.8.21	L1-36*01	J1*01	8.3.11
ZIKV-204	IgG1, κ	V2-70*01	J6*02	D4-17*01	10.7.24	K3-20*01	J4*01	7.3.9
ZIKV-218	IgG2, λ	V3-30-3*01	J4*02	D6-13*01	8.8.17	L2-11*01	J2*01	9.3.9

nd, not determined; as ambiguous results were obtained despite repeat testing.

Extended Data Table 2 | Research subjects, with time and place of infection

ZIKV strain lineage	Subject	Year infected	Country in which infection occurred
African	972	2008	Senegal
	973	2008	Sexual transmission from Subject 972*
Asian	1001	2015	Brazil
	1002	2016	Mexico
	1010	2016	Haiti
	1011	2016	Haiti
	1012	2016	Haiti
	1016	2016	Haiti

*Case was reported previously (see ref. 24).

Receptor usage dictates HIV-1 restriction by human TRIM5 α in dendritic cell subsets

Carla M. S. Ribeiro¹, Ramin Sarrami-Forooshani¹, Laurentia C. Setiawan¹, Esther M. Zijlstra-Willems¹, John L. van Hamme¹, Wikky Tigchelaar², Nicole N. van der Wel², Neeltje A. Kootstra¹, Sonja I. Gringhuis¹ & Teunis B. H. Geijtenbeek¹

The most prevalent route of HIV-1 infection is across mucosal tissues after sexual contact. Langerhans cells (LCs) belong to the subset of dendritic cells (DCs) that line the mucosal epithelia of vagina and foreskin and have the ability to sense and induce immunity to invading pathogens¹. Anatomical and functional characteristics make LCs one of the primary targets of HIV-1 infection². Notably, LCs form a protective barrier against HIV-1 infection and transmission^{3–5}. LCs restrict HIV-1 infection through the capture of HIV-1 by the C-type lectin receptor Langerin and subsequent internalization into Birbeck granules⁵. However, the underlying molecular mechanism of HIV-1 restriction in LCs remains unknown. Here we show that human E3-ubiquitin ligase tri-partite-containing motif 5 α (TRIM5 α) potently restricts HIV-1 infection of LCs but not of subepithelial DC-SIGN⁺ DCs. HIV-1 restriction by TRIM5 α was thus far considered to be reserved to non-human primate TRIM5 α orthologues^{6–9}, but our data strongly suggest that human TRIM5 α is a cell-specific restriction factor dependent on C-type lectin receptor function. Our findings highlight the importance of HIV-1 binding to Langerin for the routing of HIV-1 into the human TRIM5 α -mediated restriction pathway. TRIM5 α mediates the assembly of an autophagy-activating scaffold to Langerin, which targets HIV-1 for autophagic degradation and prevents infection of LCs. By contrast, HIV-1 binding to DC-SIGN⁺ DCs leads to disassociation of TRIM5 α from DC-SIGN, which abrogates TRIM5 α restriction. Thus, our data strongly suggest that restriction by human TRIM5 α is controlled by C-type-lectin-receptor-dependent uptake of HIV-1, dictating protection or infection of human DC subsets. Therapeutic interventions that incorporate C-type lectin receptors and autophagy-targeting strategies could thus provide cell-mediated resistance to HIV-1 in humans.

HIV-1 restriction by Langerin occurs after HIV-1 fusion but before integration of viral DNA into the host genome¹⁰. Therefore, we investigated the role of TRIM5 α , as this E3-ubiquitin ligase is a host restriction factor that restricts retroviruses after fusion by binding incoming retroviral capsid and interfering with the uncoating and reverse-transcription processes^{6,11–13}. We used both primary human LCs and human MUTZ3-derived LCs (MUTZ-LCs). Langerin on MUTZ-LCs¹⁴, as on primary LCs⁵, controls HIV-1 restriction mechanisms (Extended Data Fig. 1a–e). Notably, silencing of TRIM5 α in human LCs by RNA interference (Extended Data Fig. 2a, b, h) resulted in increased viral integration and infection with both CXCR4- and CCR5-tropic viruses, as well as increased HIV-1 transmission to activated CD4⁺ T cells (Fig. 1a–e, Extended Data Fig. 3a, b). These data strongly suggest that human TRIM5 α is a potent restriction factor for HIV-1 in LCs.

We next investigated the molecular mechanism of human TRIM5 α restriction and the interplay with autophagy machinery, as the latter has been implicated in the function of TRIM molecules¹⁵. At steady-state, human TRIM5 α associated in LCs with autophagosomal

molecule Atg16L1 and adaptor protein p62 (Extended Data Fig. 4a). Atg16L1 not only binds to ubiquitin-decorated cargos but also forms large protein complexes with Atg5 and Atg12 to elicit autophagosome biogenesis by building protein scaffolds as well as mediating expansion of autophagosomes^{16,17}. Notably, HIV-1 infection of LCs increased Atg5 recruitment to TRIM5 α -Atg16L1-p24 capsid complexes (Fig. 2a). These data suggest that human TRIM5 α in LCs is involved in the assembly of core autophagic factors into autophagosome formation upon recruitment of HIV-1 p24 capsid. Notably, HIV-1 infection of LCs increased the number of autophagosomes (bi- or multilamellar vesicles) compared to uninfected LCs (Fig. 2b, c). Furthermore, although HIV-1 infection alone did

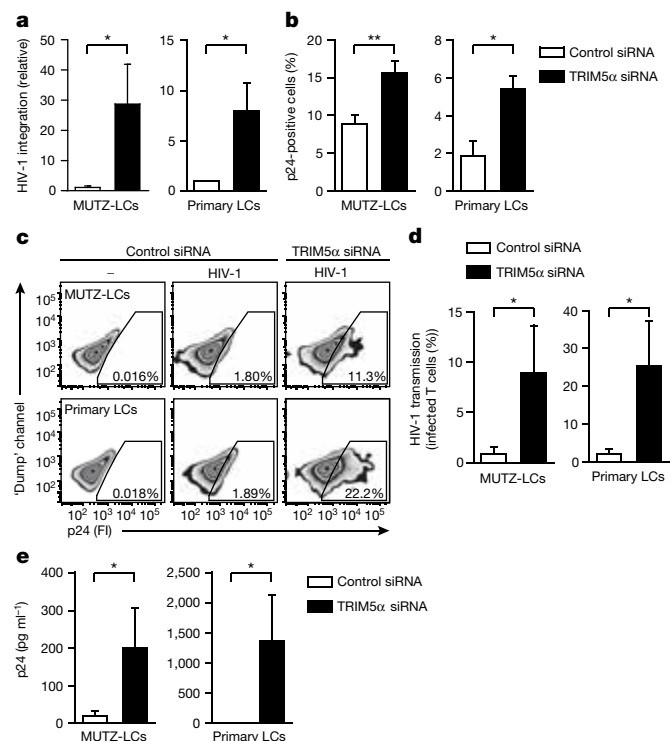


Figure 1 | Human TRIM5 α is a restriction factor for HIV-1 in LCs.

a, b, HIV-1_{NL4.3} integration (**a**) and infection (**b**) of MUTZ-LCs and primary LCs after TRIM5 α silencing, determined by Alu-PCR (**a**) and intracellular p24 staining (**b**). siRNA, small interfering RNA. **c–e**, HIV-1_{SF162} transmission by MUTZ-LCs and primary LCs after TRIM5 α silencing, determined in LC–T-cell coculture by intracellular p24 staining (**c** (representative of $n = 4$), **d**) and p24-antigen ELISA (**e**). FI, fluorescence intensity. * $P < 0.05$, ** $P < 0.01$ (two-tailed t -test). Data are mean \pm s.d. of four (**a**, MUTZ-LCs), three (**a** (primary LCs), **b**) and four (**d**, **e**) independent experiments.

¹Department of Experimental Immunology, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. ²Department of Cell Biology & Histology, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands.

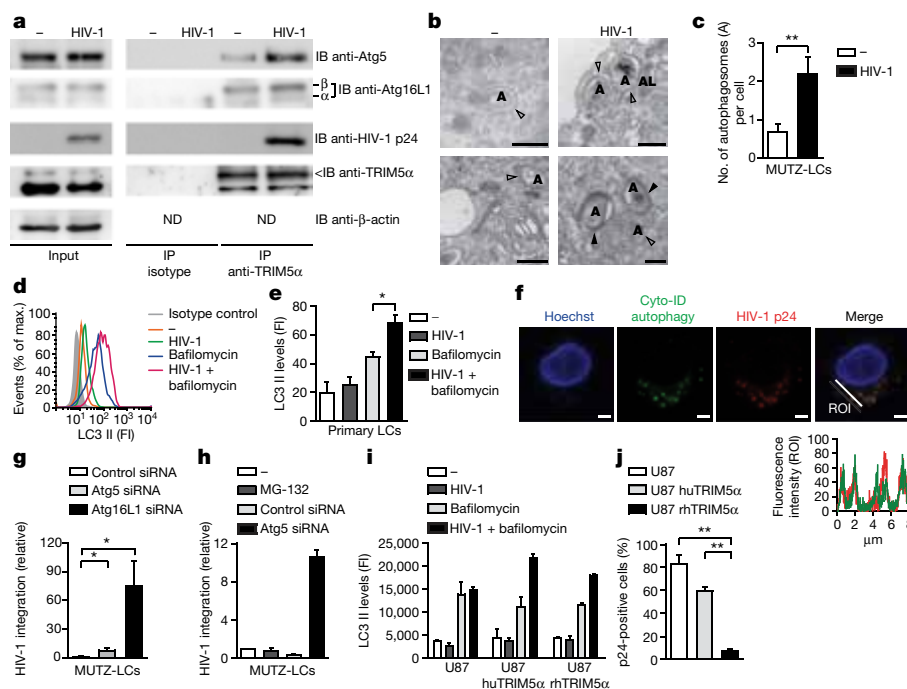


Figure 2 | Autophagy restricts HIV-1 infection of LCs. **a**, Atg5, Atg16L1, HIV-1 p24 and TRIM5 α in whole-cell lysates of MUTZ-LCs infected with HIV-1_{NL4.3} before (input) or after immunoprecipitation (IP) with TRIM5 α , determined by immunoblotting (IB); representative of $n = 3$. ND, not determined. For gel source data, see Supplementary Fig. 1. **b**, **c**, Electron microscopy analyses for bi- (empty arrowheads) and multi-lamellar autophagosomes (filled arrowheads) in HIV-1_{NL4.3}-infected MUTZ-LCs. A, autophagosomes; AL, autolysosomes. Scale bar, 500 nm. Representative of $n = 3$ (**b**); mean \pm s.d., $n = 50$ images per condition (**c**). **d**, **e**, Autophagy induction in primary LCs pre-treated with bafilomycin followed by incubation with HIV-1_{NL4.3}, determined by intracellular LC3 II levels

(representative of $n = 3$ (**d**)). **f**, Confocal microscopy analyses of primary LCs infected with HIV-1_{NL4.3}. Scale bars, 2.5 μ m. Histogram of Cyto-ID and p24 fluorescence intensities (ROI, region of interest; representative of $n = 2$). **g**, **h**, HIV-1_{NL4.3} integration into MUTZ-LCs after Atg5 or Atg16L1 silencing or after pre-treatment with MG-132, determined by Alu-PCR. **i**, Autophagy induction in U87 cells or transduced with human or rhesus TRIM5 α pre-treated with bafilomycin followed by incubation with HIV-1_{SF162}, determined by intracellular LC3 II levels ($n = 2$). **j**, HIV-1_{SF162} infection of U87 transfectants, determined by intracellular p24 staining. * $P < 0.05$, ** $P < 0.01$ (two-tailed t -test). Data are mean \pm s.d. of three (**e**, **h**, **j**) and four (**g**) independent experiments.

not affect microtubule-associated protein light chain 3 II (LC3 II) levels compared to uninfected cells, these levels were increased after pre-treatment with lysosomal inhibitor bafilomycin of HIV-1-infected cells (Fig. 2d, e, Extended Data Fig. 4b). These data indicate that HIV-1 increases autophagic flux in LCs, rather than blocking autophagic maturation.

Therefore, we investigated whether autophagy restricts HIV-1 infection in LCs. HIV-1 p24 capsid co-immunoprecipitated with autophagic molecules (Fig. 2a) and we observed targeting of HIV-1 p24 capsid into autophagic vesicles in LCs (Fig. 2f). Notably, silencing of Atg5 or Atg16L1 increased both HIV-1 integration and infection of LCs (Fig. 2g, h, Extended Data Figs 2a, i, j, 4c). Similarly, an increase of HIV-1 integration was observed after silencing Atg13 or FIP200 (Extended Data Fig. 4d), suggesting that the ULK1-dependent autophagy pathway prevents infection of LCs. Enhancing autophagy-mediated lysis by rapamycin decreased HIV-1 integration in primary LCs and led to degradation of HIV-1 p24 in primary LCs (Fig. 3a–c, g). Collectively, these data strongly suggest that, upon viral fusion, HIV-1 capsids are targeted into autophagosomes for lysosomal degradation. Furthermore, silencing of Atg5 in TRIM5 α -silenced LCs did not further affect either HIV-1 integration or infection, supporting the notion that the mechanism of TRIM5 α restriction is dependent on Atg5 function (Extended Data Fig. 2d, e, 4e, f). The role for TRIM5 α in autophagy activation was further supported by our data showing that HIV-1 infection of CD4⁺CCR5⁺ U87 cells that overexpressed either human or rhesus TRIM5 α increased Atg5 recruitment to Atg16L1–TRIM5 α complexes (Extended Data Fig. 5a), and increased LC3 II levels in the presence of bafilomycin (Fig. 2i, Extended Data Fig. 5b). In line with previous reports^{6,7,11,15}, rhesus but not human TRIM5 α strongly restricted HIV-1 infection (Fig. 2j). Thus, human TRIM5 α is unable to restrict

HIV-1 infection in U87 cells despite induction of autophagy upon HIV-1 infection. Therefore, we hypothesized that LC-specific uptake through Langerin might drive efficient human TRIM5 α restriction. As vesicular stomatitis virus-G glycoprotein (VSV-G)-pseudotyped viruses do not interact with Langerin¹⁸ and infect cells through endocytosis-mediated uptake independently of CD4 and CCR5, we silenced endogenous TRIM5 α in LCs and infected them with HIV-1 isolate NL4.3 or VSV-G-pseudotyped NL4.3(Δ Env) HIV-1. Silencing of human TRIM5 α in LCs increased integration of HIV-1, but notably not of VSV-G-pseudotyped HIV-1 (Fig. 3d). These findings strongly suggest that human TRIM5 α restriction depends on the virus uptake route through Langerin. Therefore, we investigated whether Langerin is part of the TRIM5 α –autophagy complex. Atg16L1 and TRIM5 α co-immunoprecipitated with Langerin in LCs at both steady state and after HIV-1 exposure (Fig. 3e). Confocal microscopy confirmed the partial colocalization of human TRIM5 α and Langerin in primary LCs (Fig. 3f). These data suggest that Langerin couples HIV-1 routing to human TRIM5 α -mediated autophagic restriction. Notably, although rapamycin, which induces autophagy, alone did not affect Langerin expression, HIV-1 infection strongly decreased Langerin levels independently of rapamycin (Fig. 3g). We could also detect the presence of Birbeck granules within autophagosomes in LCs by electron microscopy (Fig. 3h, i). These findings suggest that autophagy machinery intersects with Langerin internalization pathway, and vesicles as well as whole Birbeck granules containing Langerin–HIV-1 capsid complexes are turned over by autophagy upon viral fusion in LCs.

Notably, ectopic expression of Langerin in CD4⁺CCR5⁺ U87 cells strongly restricted HIV-1 integration and infection (Fig. 3j–l), and silencing of either TRIM5 α or Atg16L1 in these Langerin-expressing

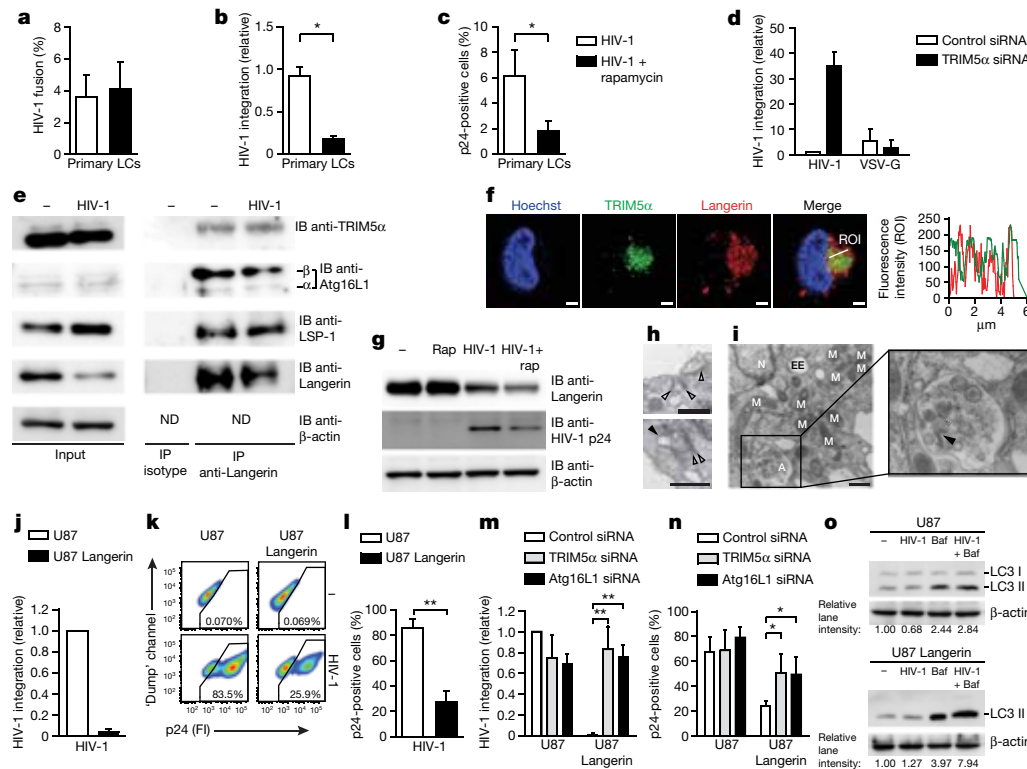


Figure 3 | HIV-1 uptake by Langerin drives human TRIM5 α restriction.

a–c, Primary LCs pre-treated with rapamycin, HIV-1_{NL4.3}-BlaM-Vpr fusion, determined by β -lactamase-Vpr (BlaM-Vpr) assay (**a**), HIV-1_{NL4.3} integration, determined by Alu-PCR (**b**), HIV-1_{NL4.3} infection, determined by intracellular p24 staining (**c**). **d**, HIV-1_{NL4.3} or VSV-G-pseudotyped HIV-1_{NL4.3} integration into MUTZ-LCs after TRIM5 α silencing, determined by Alu-PCR ($n = 2$). **e**, TRIM5 α , Atg16L1, LSP-1 and Langerin in whole-cell lysates of MUTZ-LCs infected with HIV-1_{NL4.3} before (input) or after immunoprecipitation with Langerin, determined by immunoblotting (representative of $n = 3$). For gel source data, see Supplementary Fig. 1. **f**, Confocal microscopy analyses of primary LCs. Scale bars, 2.5 μ m. Histogram of Langerin and TRIM5 α fluorescence intensities (ROI, region of interest). **g**, Langerin and HIV-p24 in primary LCs pre-treated with rapamycin followed by incubation with HIV-1_{NL4.3}, determined by immunoblotting. **h**, **i**, Electron microscopy analyses

for tubular-shaped (empty arrowheads) and racket-shaped (filled arrowheads) Birbeck granules in MUTZ-LCs after incubation with HIV-1_{NL4.3}. Scale bars, 500 nm (**i**; original magnification, $2.5 \times$ (inset)). EE, early endosomes; M, mitochondria; N, nucleus. Representative of $n = 2$ (**f–i**). **j–l**, HIV-1_{NL4.3}-BaL integration (**j**) or infection (**k**, **l**) of U87 or Langerin⁺ U87 cells, determined by Alu-PCR (**j**) and intracellular p24 staining (**k** (representative of $n = 4$), **l**). **m**, **n**, HIV-1_{NL4.3}-BaL integration (**m**) and infection (**n**) of U87 transfectants after Atg16L1 or TRIM5 α silencing, determined by Alu-PCR (**m**) and intracellular p24 staining (**n**). **o**, Autophagy induction in U87 transfectants pre-treated with bafilomycin followed by incubation with HIV-1_{SF162}, determined by immunoblotting for LC3. Relative abundance of LC3II determined by normalizing to β -actin, representative of $n = 2$. * $P < 0.05$, ** $P < 0.01$ (two-tailed t -test). Data are mean \pm s.d. of four (**a**, **l**, **n**) and three (**b**, **c**, **j**, **m**) independent experiments.

U87 cells abrogated restriction (Extended Data Fig. 2f, Fig. 3m, n). Furthermore, increased levels of the TRIM5 α -Atg5 complexes co-immunoprecipitated with Atg16L1 and correlated with increased levels of LC3 II upon HIV-1 infection in Langerin-expressing U87 cells, but not in the U87 parental cells (Extended Data Fig. 6a, Fig. 3o). Silencing of human TRIM5 α decreased HIV-induced autophagy in Langerin-expressing U87 cells (Extended Data Fig. 6b), which supports the role for TRIM5 α in mediating Langerin-dependent autophagic restriction of HIV-1.

We next investigated whether other HIV-1-binding C-type lectin receptors, such as DC-SIGN, can also recruit TRIM5 α machinery. DC-SIGN, in contrast to Langerin on LCs, facilitates HIV-1 infection of and transmission by DCs^{19,20}. Silencing of TRIM5 α in DC-SIGN⁺ DCs (Extended Data Fig. 2c) affected neither HIV-1 integration nor infection levels of both CXCR4- and CCR5-tropic viruses (Fig. 4a–c, Extended Data Fig. 3c, d). Furthermore, HIV-1 downregulated autophagy in DCs (Fig. 4d), as reported previously²¹. Notably, although TRIM5 α co-immunoprecipitated with an antibody against DC-SIGN at steady-state conditions, HIV-1 infection in DCs led to dissociation of TRIM5 α from the cytoplasmic domain of DC-SIGN (Fig. 4e). These data underscore that human TRIM5 α is a cell-specific restriction factor and support the hypothesis that other C-type lectin receptors also recruit TRIM5 α , but only Langerin seems to have the ability to restrict HIV-1 infection through TRIM5 α .

To further identify the molecular determinants of TRIM5 α restriction, we transduced U87 cells with Langerin(W264R) mutant, which contains a naturally occurring polymorphism within the binding pocket of the extracellular carbohydrate-recognition domain of Langerin and is unable to bind sugars²². The Langerin(W264R) mutant did not bind HIV-1 and was unable to restrict HIV-1 infection in U87 cells (Fig. 4f–h), demonstrating that binding to the carbohydrate-recognition domain of Langerin is required for efficient TRIM5 α restriction. We have previously shown that differential binding of specific signalling molecules to the cytoplasmic domain of DC-SIGN through adaptor protein LSP-1 dictates intracellular signalling²³. Similar to DC-SIGN (Fig. 4e), Langerin also interacted with LSP-1 (Fig. 3e)²⁴. LSP-1 forms a complex with Langerin, TRIM5 α and the autophagosomal molecule Atg16L1 in LCs (Fig. 3e). Immunoprecipitation analyses confirmed the interaction between TRIM5 α and LSP-1 in LCs (Extended Data Fig. 6c). Silencing of LSP-1, but not Atg16L1, decreased Langerin binding to TRIM5 α (Fig. 4i, Extended Data Fig. 6d). Remarkably, HIV-1 binding to the extracellular carbohydrate-recognition domain of either Langerin or DC-SIGN dictates TRIM5 α assembly to the intracellular domain of the C-type lectin-receptor through LSP-1, leading to HIV-1 restriction or infection, respectively. The LC-specific restriction mechanism identified in our study not only highlights the natural barrier function of the mucosa to HIV-1, but also outlines a novel receptor-controlled TRIM5 α restriction mechanism.

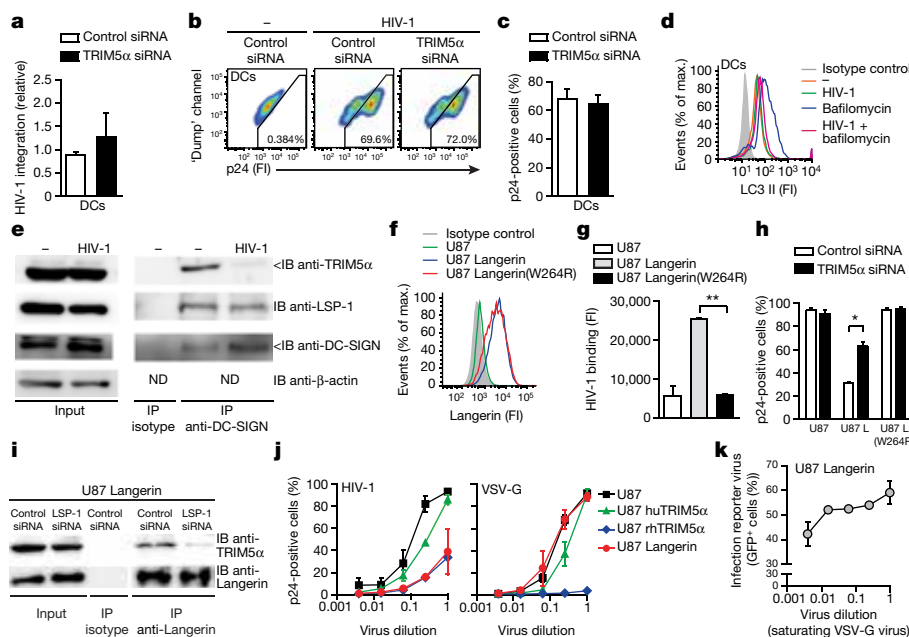


Figure 4 | Human TRIM5α is a cell-specific restriction factor for HIV-1.

a–c, HIV-1_{NL4.3-BaL} integration (**a**) and infection (**b**, **c**) of DCs after TRIM5α silencing, determined by Alu-PCR (**a**) and intracellular p24 staining (**b** (representative of $n = 6$), **c**). **d**, Autophagy induction in DCs pre-treated with bafilomycin followed by incubation with HIV-1_{NL4.3-BaL}, determined by intracellular LC3 II levels (representative of $n = 2$). **e**, TRIM5α, LSP-1 and DC-SIGN in whole-cell lysates of DCs infected with HIV-1_{NL4.3-BaL} before (input) or after immunoprecipitation together with DC-SIGN, determined by immunoblotting (representative of $n = 3$). For gel source data, see Supplementary Fig. 1. **f**, Langerin expression in U87 parental cells or transduced with either Langerin or Langerin(W264R) mutant, determined by flow cytometry, representative

of $n = 3$. **g**, HIV-1 binding to U87 transfectants, determined by gp120 beads-binding assay. **h**, HIV-1_{NL4.3-BaL} infection of U87 transfectants after TRIM5α silencing, determined by intracellular p24 staining. **i**, TRIM5α and Langerin in whole-cell lysates of U87 Langerin transfectant after LSP-1 silencing before (input) or after immunoprecipitation together with Langerin, determined by immunoblotting, representative of $n = 2$. **j**, HIV-1_{NL4.3-BaL} or VSV-G-pseudotyped HIV-1 infection of U87 transfectants, determined by intracellular p24 staining. **k**, Abrogation of restriction in U87 Langerin transfectant after pre-incubation with increasing doses of VSV-G-pseudotyped particles followed by infection with HIV-1_{NL4.3-GFP-BaL} reporter virus (**j**, **k**; $n = 2$). * $P < 0.05$, ** $P < 0.01$ (two-tailed t -test). Data are mean \pm s.d. of three (**a**, **g**, **h**) and six (**c**) independent experiments.

We next compared the restriction mechanism of human TRIM5α in Langerin-expressing U87 cells with that described for rhesus TRIM5α; the two mechanisms seem to differ as silencing of Atg16L1 only relieved rhesus TRIM5α-induced restriction twofold (Extended Data Figs 2g, 5c). In accordance with a previous study²⁵, proteasome inhibition with MG-132 rescued reverse transcription, but not infection, of U87 cells transduced with rhesus TRIM5α (Extended Data Fig. 7a, b). By contrast, MG-132 did not abrogate HIV-1 restriction in MUTZ-LCs or in Langerin-expressing U87 cells (Fig. 2h, Extended Data Fig. 7c, d). These data further support that restriction by the Langerin-mediated TRIM5α mechanism depends on autophagy machinery and differs from the proteasome-dependent restriction by rhesus TRIM5α. These findings challenge the current hypothesis that human TRIM5α cannot restrict HIV-1 infection^{6,7,26,27}, although they suggest a mechanism that is distinct from that of rhesus TRIM5α. We next investigated whether the viral envelope affects restriction. Rhesus TRIM5α restricts both HIV-1- and VSV-G-pseudotyped viruses (Fig. 4j), in accordance with previous reports^{27,28}. Notably, our data show that the Langerin-controlled human TRIM5α mechanism restricts infection of HIV-1 at the same magnitude as rhesus TRIM5α, but does not restrict infection of VSV-G-pseudotyped virus (which bypasses Langerin uptake). These findings indicate that TRIM5α in human cells, in contrast to rhesus TRIM5α⁶, depends on the virus uptake route by Langerin. Furthermore, our data show that TRIM5α restriction in U87 cells transduced with Langerin is saturated by increasing doses of viral capsids, thus rendering cells permissive to infection by a HIV-1 reporter virus (Fig. 4k), as previously shown for rhesus TRIM5α^{29,30}. As VSV-G-pseudotyped virus induced autophagy at saturating conditions (Extended Data Fig. 6e), these data strongly suggest that autophagy induction alone does not mediate human TRIM5α-dependent restriction of HIV-1 (Fig. 2i, j, 4k, Extended Data Fig. 6e),

but that it requires the formation of a complex between HIV-1 p24 capsid, human TRIM5α and autophagy molecules for restriction (Figs 2a, 4k, Extended Data Fig. 8). Further studies are required to investigate whether direct interaction between human TRIM5α and HIV-1 p24 capsid is required for the Langerin-mediated TRIM5α restriction mechanism. Our data demonstrate that restriction by human TRIM5α is two-tiered; TRIM5α mediates assembly of the autophagy-activating complexes and at the same time requires HIV-1 uptake by Langerin for efficient routing of the HIV-1 capsid into TRIM5α-dependent autophagic scaffolds. Our study establishes that human TRIM5α is a cell-specific restriction factor for HIV-1 and underscores the importance of selective HIV-1 uptake mechanisms and assembly of molecular determinants in driving human TRIM5α-mediated restriction. Novel TRIM5α-based therapies in combination with strategies targeting C-type lectin receptors, LSP-1 or autophagy could thus represent an important alternative to current antiretroviral therapy in acute retroviral exposure.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 March; accepted 20 October 2016.

Published online 7 December 2016.

1. Banchereau, J. *et al.* Immunobiology of dendritic cells. *Annu. Rev. Immunol.* **18**, 767–811 (2000).
2. Hladik, F. *et al.* Initial events in establishing vaginal entry and infection by human immunodeficiency virus type-1. *Immunity* **26**, 257–270 (2007).
3. Ribeiro, C. M. S., Sarrami-Forooshani, R. & Geijtenbeek, T. B. H. HIV-1 border patrols: Langerhans cells control antiviral responses and viral transmission. *Future Virol.* **10**, 1231–1243 (2015).
4. Sarrami-Forooshani, R. *et al.* Human immature Langerhans cells restrict CXCR4-using HIV-1 transmission. *Retrovirology* **11**, 52 (2014).
5. de Witte, L. *et al.* Langerin is a natural barrier to HIV-1 transmission by Langerhans cells. *Nat. Med.* **13**, 367–371 (2007).

6. Stremlau, M. *et al.* The cytoplasmic body component TRIM5 α restricts HIV-1 infection in Old World monkeys. *Nature* **427**, 848–853 (2004).
7. Sawyer, S. L., Wu, L. I., Emerman, M. & Malik, H. S. Positive selection of primate TRIM5 α identifies a critical species-specific retroviral restriction domain. *Proc. Natl Acad. Sci. USA* **102**, 2832–2837 (2005).
8. Song, B. *et al.* Retrovirus restriction by TRIM5 α variants from Old World and New World primates. *J. Virol.* **79**, 3930–3937 (2005).
9. Sayah, D. M., Sokolskaja, E., Berthou, L. & Luban, J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**, 569–573 (2004).
10. van den Berg, L. M. *et al.* Caveolin-1 mediated uptake via Langerin restricts HIV-1 infection in human Langerhans cells. *Retrovirology* **11**, 123 (2014).
11. Stremlau, M. *et al.* Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5 α restriction factor. *Proc. Natl Acad. Sci. USA* **103**, 5514–5519 (2006).
12. Yap, M. W., Nisole, S. & Stoye, J. P. A single amino acid change in the SPRY domain of human Trim5 α leads to HIV-1 restriction. *Curr. Biol.* **15**, 73–78 (2005).
13. Ganser-Pornillos, B. K. *et al.* Hexagonal assembly of a restricting TRIM5 α protein. *Proc. Natl Acad. Sci. USA* **108**, 534–539 (2011).
14. de Jong, M. A. *et al.* Mutz-3-derived Langerhans cells are a model to study HIV-1 transmission and potential inhibitors. *J. Leukoc. Biol.* **87**, 637–643 (2010).
15. Mandell, M. A. *et al.* TRIM proteins regulate autophagy and can target autophagic substrates by direct recognition. *Dev. Cell* **30**, 394–409 (2014).
16. Fujita, N. *et al.* Recruitment of the autophagic machinery to endosomes during infection is mediated by ubiquitin. *J. Cell Biol.* **203**, 115–128 (2013).
17. Moreau, K., Ravikumar, B., Renna, M., Puri, C. & Rubinshtein, D. C. Autophagosome precursor maturation requires homotypic fusion. *Cell* **146**, 303–317 (2011).
18. Gramberg, T. *et al.* Interactions of LSECtin and DC-SIGN/DC-SIGNR with viral ligands: differential pH dependence, internalization and virion binding. *Virology* **373**, 189–201 (2008).
19. Geijtenbeek, T. B. *et al.* DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances *trans*-infection of T cells. *Cell* **100**, 587–597 (2000).
20. Gringhuis, S. I. *et al.* HIV-1 exploits innate signaling by TLR8 and DC-SIGN for productive infection of dendritic cells. *Nat. Immunol.* **11**, 419–426 (2010).
21. Blanchet, F. P. *et al.* Human immunodeficiency virus-1 inhibition of immunoamphisomes in dendritic cells impairs early innate and adaptive immune responses. *Immunity* **32**, 654–669 (2010).
22. Ward, E. M., Stambach, N. S., Drickamer, K. & Taylor, M. E. Polymorphisms in human Langerin affect stability and sugar binding activity. *J. Biol. Chem.* **281**, 15450–15456 (2006).
23. Gringhuis, S. I., den Dunnen, J., Litjens, M., van der Vlist, M. & Geijtenbeek, T. B. Carbohydrate-specific signaling through the DC-SIGN signalosome tailors immunity to *Mycobacterium tuberculosis*, HIV-1 and *Helicobacter pylori*. *Nat. Immunol.* **10**, 1081–1088 (2009).
24. Smith, A. L. *et al.* Leukocyte-specific protein 1 interacts with DC-SIGN and mediates transport of HIV to the proteasome in dendritic cells. *J. Exp. Med.* **204**, 421–430 (2007).
25. Wu, X., Anderson, J. L., Campbell, E. M., Joseph, A. M. & Hope, T. J. Proteasome inhibitors uncouple rhesus TRIM5 α restriction of HIV-1 reverse transcription and infection. *Proc. Natl Acad. Sci. USA* **103**, 7465–7470 (2006).
26. Perez-Caballero, D., Hatzioannou, T., Yang, A., Cowan, S. & Bieniasz, P. D. Human tripartite motif 5 α domains responsible for retrovirus restriction activity and specificity. *J. Virol.* **79**, 8969–8978 (2005).
27. Stremlau, M., Perron, M., Welikala, S. & Sodroski, J. Species-specific variation in the B30.2(SPRY) domain of TRIM5 α determines the potency of human immunodeficiency virus restriction. *J. Virol.* **79**, 3139–3145 (2005).
28. Song, B. *et al.* The B30.2(SPRY) domain of the retroviral restriction factor TRIM5 α exhibits lineage-specific length and sequence variation in primates. *J. Virol.* **79**, 6111–6121 (2005).
29. Shi, J. & Aiken, C. Saturation of TRIM5 α -mediated restriction of HIV-1 infection depends on the stability of the incoming viral capsid. *Virology* **350**, 493–500 (2006).
30. Kootstra, N. A., Munk, C., Tonnu, N., Landau, N. R. & Verma, I. M. Abrogation of postentry restriction of HIV-1-based lentiviral vector transduction in simian cells. *Proc. Natl Acad. Sci. USA* **100**, 1298–1303 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to the members of the Host Defense group and Laboratory for Viral Immune Pathogenesis (Department of Experimental Immunology, Academic Medical Center, Amsterdam, The Netherlands) for their input and D. Picavet (van Leeuwenhoek Centrum for Advanced Microscopy, Academic Medical Center, Amsterdam, The Netherlands) for technical assistance during confocal experiments. We wish to thank the Boerhaave Medical Centre (Amsterdam, The Netherlands) and A. Knottenbelt (Flevoclinic, Almere, The Netherlands) for the provision of human skin tissues. This work was supported by the Dutch Scientific Organization NWO (VENI 863.13.025 and VICI 918.10.619), Aids Fonds (2010038) and European Research Council (Advanced grant 670424).

Author Contributions C.M.S.R. designed, performed and interpreted most experiments and prepared the manuscript; R.S.F. assisted with the lentiviral transductions and the confocal experiments; L.C.S. assisted with culturing the CD4⁺CCR5⁺ U87 cell lines and with the lentiviral transductions; E.M.Z.W. cultured MUTZ-LCs and helped with immunoblotting; J.L.v.H. assisted with primary cell isolation and silencing experiments; W.T. and N.N.v.d.W. performed the EM microscopy; N.A.K. and S.I.G. helped prepare the manuscript and T.B.H.G. supervised all aspects of the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.B.H.G. (t.b.geijtenbeek@amc.uva.nl) or C.M.S.R. (c.m.ribeiro@amc.uva.nl).

Reviewer Information *Nature* thanks J. Luban, C. Munz and G. Towers for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Donors, cells and inhibitors. Human skin tissue was obtained from healthy donors undergoing corrective breast or abdominal surgery after informed consent in accordance with our institutional guidelines. This study was approved by the Medical Ethics Review Committee of the Academic Medical Center. Split-skin grafts of 0.3 mm in thickness were obtained using a dermatome (Zimmer). After incubation with Dispase II (1 U ml^{-1} , Roche Diagnostics), epidermal sheets were separated from the dermis and cultured in in Iscoves Modified Dulbeccos's Medium (IMDM, Thermo Fischer Scientific) supplemented with 10% FCS, gentamycin ($20 \mu\text{g ml}^{-1}$, Centrafarm), penicillin/streptomycin (100 U ml^{-1} and $10 \mu\text{g ml}^{-1}$, respectively; Invitrogen). Further LC purification was performed using a Ficoll gradient (Axis-shield) and CD1a microbeads (Miltenyl Biotec) as described before^{4,10}. Isolated LCs were routinely 90% pure and expressed high levels of Langerin and CD1a. MUTZ-LCs were differentiated from CD34⁺ human AML cell line MUTZ3 progenitors in the presence of GM-CSF (100 ng ml^{-1} , Invitrogen), TGF- β (10 ng ml^{-1} , R&D) and TNF- α (2.5 ng ml^{-1} , R&D) and cultured as described before¹⁴. Immature DCs were differentiated from monocytes, isolated from buffy coats of healthy volunteer blood donors (Sanquin, The Netherlands), in the presence of IL-4 (500 U ml^{-1} , Invitrogen) and GM-CSF (800 U ml^{-1} , Invitrogen) and used at day 6 or 7 as previously described²⁰. CD4⁺ T cells were obtained from peripheral blood mononuclear cells (PBMCs) activated with phytohaemagglutinin (1 mg ml^{-1} ; L2769, Sigma Aldrich) for 3 days, enriched for CD4⁺ T cells by negative selection using MACS beads (130-096-533, Miltenyi) and cultured overnight with IL-2 (20 U ml^{-1} ; 130-097-745, Miltenyi) as described before⁵. The following inhibitors were used: rapamycin (mTOR inhibitor; tlr-rap, Invivogen), bafilomycin A1 (V-ATPase inhibitor; tlr-baf1; Invivogen) and MG-132 (proteasome inhibitor; 474790; Calbiochem).

Plasmids and cell lines. All cell lines were obtained from ATCC and tested negative for mycoplasma contamination, determined in 3-day-old cell cultures by PCR. Langerin and Langerin mutant W264R expression plasmid pcDNA3.1 were obtained from Life Technologies and subcloned into lentiviral construct pWPXLd (Addgene). HIV-1-based lentiviruses were produced by co-transfection of 293T cells with the lentiviral vector construct, the packaging construct (psPAX2, Addgene) and vesicular stomatitis virus glycoprotein envelope (pMD2.G, Addgene) as described previously³¹. U87 cell lines stably expressing CD4 and wild-type CCR5 co-receptor (obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: U87 CD4⁺CCR5⁺ cells from H. K. Deng and D. R. Littman³²) were transduced with HIV-1-based lentiviruses expressing sequences coding human TRIM5 α ³³, rhesus TRIM5 α ³³, wild-type Langerin or Langerin(W264R).

Viruses, HIV-1 infection and transmission. NL4.3, NL4.3-BaL, SF162, NL4.3eGFP-BaL, NL4.3-BlaM-Vpr and VSV-G-pseudotyped NL4.3(Δ Env) HIV-1 were generated as described¹⁰. All produced viruses were quantified by p24 ELISA (Perkin Elmer Life Sciences) and titrated using the indicator cells TZM-BL. Primary LCs and MUTZ-LCs were infected with a multiplicity of infection of 0.2–0.4 and HIV-1 infection was assessed by flow cytometry at day 7 after infection by intracellular p24 staining. Double staining with CD1a (LCs marker; HI149-APC; BD Pharmingen) and p24 (KC57-RD1-PE; Beckman Coulter) was used to discriminate the percentage of CD1a⁺p24⁺ infected LCs. CD4⁺CCR5⁺ U87 parental or transduced cells were infected at a multiplicity of infection of 0.1–0.2 and HIV-1 infection was assessed at day 3 after infection by intracellular p24 staining or GFP expression. For analysis of transmission of HIV-1 to T cells, LCs were stringently washed 3 days after infection followed by co-culture with activated allogeneic CD4⁺ T cells for 3 days. Triple staining with CD1a (LCs marker), CD3 (T cells marker; 552851-PerCP, BD Pharmingen) and p24 was used to discriminate the percentage of CD3⁺CD1a⁺p24⁺ infected T cells. HIV-1 infection and transmission was assessed by FACSCanto II flow cytometer (BD Biosciences) and data analysis was carried out with FlowJo software (Treestar). HIV-1 production was determined by a p24 antigen ELISA in culture supernatants (ZeptoMetrix).

RNA isolation and quantitative real-time PCR. mRNA was isolated with an mRNA Capture kit (Roche) and cDNA was synthesized with a reverse-transcriptase kit (Promega). For real-time PCR analysis, PCR amplification was performed in the presence of SYBR green in a 7500 Fast Realtime PCR System (ABI). Specific primers were designed with Primer Express 2.0 (Applied Biosystems; Extended Data Table 1). The cycling threshold (C_t) value is defined as the number of PCR cycles in which the fluorescence signal exceeds the detection threshold value. For each sample, the normalized amount of target mRNA (N_t) was calculated from the C_t values obtained for both target and housekeeping (GAPDH, primary LCs, DCs and U87 cells lines; β -actin, MUTZ-LCs) mRNA with the equation $N_t = 2^{C_t(\text{control}) - C_t(\text{target})}$. For relative mRNA expression, control siRNA sample was set at 1 within the experiment and for each donor.

HIV-1 integration Alu-PCR assay. A two-step Alu-long terminal repeat (LTR) PCR was used to quantify the integrated HIV-1 DNA in infected cells as previously described²⁰. Total cell DNA was isolated at 16 h after infection (multiplicity of infection of 0.4) with a QIAamp blood isolation kit (Qiagen). In the first round of PCR, the DNA sequence between HIV-1 LTR (LTR R region, extended with a marker region at the 5' end) and the nearest Alu repeat was amplified (primer sequences, Extended Data Table 1). The second round was nested quantitative real-time PCR of the first-round PCR products using primers annealing to the aforementioned marker region in combination with another HIV-1-specific primer (LTR U5 region) by real-time quantitative PCR. Two different dilutions of the PCR products from the first-round of PCR were assayed to ensure that PCR inhibitors were absent. For monitoring the signal contributed by unintegrated HIV-1 DNA, the first-round PCR was also performed using the HIV-1-specific primer (LTR R region) only. HIV-1 integration was normalized relative to GAPDH DNA levels. For relative HIV-1 integration, control siRNA-infected cells (total signal; Supplementary Table 1) was set as 1 for one experiment or for each donor.

HIV-1 fusion assay. A BlaM-Vpr-based assay was used to quantify fusion of HIV-1 to the host membrane in infected LCs as previously described¹⁰. LCs were infected with NL4.3-BlaM-Vpr for 2 h and then loaded with CCF2/AM (1 mM, LiveBlaZer FRET-B/G Loading Kit, Life technologies) in serum-free IMDM medium for 1 h at 25 °C. After washing, BlaM reaction was allowed to develop for 16 h at 22 °C in IMDM supplemented with 10% FCS and 2.5 mM anion transport inhibitor probenecid (Sigma Pharmaceuticals). HIV-1 fusion was determined by monitoring the changes in fluorescence of CCF2/AM dye, which reflect the presence of BlaM-Vpr into the cytoplasm of target cells upon viral fusion. The shift from green emission fluorescence (500 nm) to blue emission fluorescence (450 nm) of CCF2/AM dye was assessed by flow cytometer LSRFortessa (BD Biosciences) and data analysis was carried out with FlowJo software. Percentages of blue fluorescent CCF2/AM⁺ cells are depicted as percentage of HIV-1 fusion.

HIV-1 binding assay. A fluorescent bead adhesion assay was used to examine the ability of HIV-1 gp120-coated fluorescent beads to bind Langerin in CD4⁺CCR5⁺ U87 transfectants as previously described⁵. Binding was measured by FACSCanto II flow cytometer and data analysis was carried out with FlowJo software.

RNA interference. Skin LCs and DCs were transfected with 50 nm siRNA with the transfection reagent DF4 (Dharmacon) whereas MUTZ-LCs, CD4⁺CCR5⁺ U87 parental or transduced cells were transfected with transfection reagent DF1 (Dharmacon) and were used for experiments 48–72 h after transfection. The siRNA (SMARTpool; Dharmacon) were specific for Atg5, (M-004374-04), Atg16L1 (M-021033), LSP-1 (M-012640-00), TRIM5 α (M-007100-00) and non-targeting siRNA (D-001206-13) served as control. Langerin was silenced in MUTZ-LCs by electroporation with Neon Transfection System (ThermoFischer Scientific) using siRNA Langerin (10 μM siRNA, M-013059-01, SMARTpool; Dharmacon). Silencing of the aforementioned targets was verified by real-time PCR, flow cytometer and immunoblotting (Extended Data Figs 1d, e, 2a–k).

Intracellular staining of LC3 II. Cells were pre-treated with bafilomycin A1 for 2 h or left untreated followed by incubation with HIV-1 for 16 h. Quantification of intracellular LC3 II levels by saponin extraction was performed as described before^{34,35}. LCs were washed in PBS and permeabilized with 0.05% saponin in PBS. Cells were incubated at 4 °C for 30 min with mouse anti-LC3 primary antibody (M152-3; MBL International) or with mouse anti-IgG1 isotype control (MOPC-21; BD Pharmingen) followed by incubation with Alexa Fluor 488-conjugated goat-anti mouse IgG1 antibody (A-21121, Life Technologies) in saponin buffer. Intracellular LC3 II levels were assessed by FACSScan or FACSCanto II flow cytometers (BD Biosciences) and data analysis was carried out with FlowJo.

Immunoblotting for LC3. Cells were pre-treated with bafilomycin for 2 h or left untreated followed by incubation with HIV-1 for 4 h. Quantification of intracellular LC3 II levels by saponin extraction was performed as described before³⁵. Whole-cell extracts were prepared using RIPA lysis buffer supplemented with protease inhibitors (9806; Cell Signalling). 20–30 μg of extract were resolved by SDS-PAGE (15%) and immunoblotted with LC3 (2G6; Nanotools) and β -actin (sc-81178; Santa Cruz) antibodies, followed by incubation with HRP-conjugated secondary rabbit-anti-mouse antibody (P0161; Dako) and luminol-based enhanced chemiluminescence (ECL) detection (34075; Thermo Scientific). For gel source data, see Supplementary Fig. 1.

Electron microscopy. MUTZ-LCs (2×10^6) were incubated for 16 h with HIV-1 NL4.3 (multiplicity of infection, 0.5) or left untreated as a control, fixed in 4% paraformaldehyde and 1% glutaraldehyde in sodium cacodylate buffer for 10 min at room temperature followed by 24 h at 4 °C. After fixation, cells were collected by centrifugation and the pellet was washed in sodium cacodylate buffer. Cells were post-fixed for 1 h at 4 °C (1% osmium tetroxide, 0.8% potassium ferrocyanide in the same buffer), contrasted in 0.5% uranyl acetate, dehydrated in a graded ethanol series and embedded in epon LX112. Ultrathin sections were stained with uranylacetate/lead citrate and examined with a FEI Tecnai-12 transmission electron

microscope. Numbers of autophagosomes per cell was determined in 50 cells for each condition counted by two independent researchers.

Confocal microscopy. LCs were left to adhere onto poly-L-lysine coated slides. Cells were fixed in 4% paraformaldehyde and permeabilized with PBS/0.1% saponin/1% BSA/1 mM Hepes. Cells were stained with anti-Langerin (AF2088; R&D Systems) and TRIM5 α (ab109709; Abcam) antibodies followed by Alexa Fluor 647-conjugated anti-goat (A-21447; Life Technologies) and Alexa Fluor 488-conjugated anti-rabbit (A-21206; Life Technologies). For detection of autophagic vesicles, LCs were pre-loaded with the Cyto-ID Green detection autophagy reagent (ENZ-51031; Enzo Life Sciences), which was previously shown to specifically stain autophagic vesicles³⁶ before adherence to microscope slides and stained with p24 (KC57-RD1-PE; Beckman Coulter) followed by Alexa-Fluor-546-conjugated anti-mouse (A-11003; Life Technologies). Nuclei were counterstained with Hoechst (10 μ g ml⁻¹; Molecular Probes). Single plane images were obtained by Leica TCS SP-8 X confocal microscope and data analysis was carried out with Leica LAS AF Lite (Leica Microsystems).

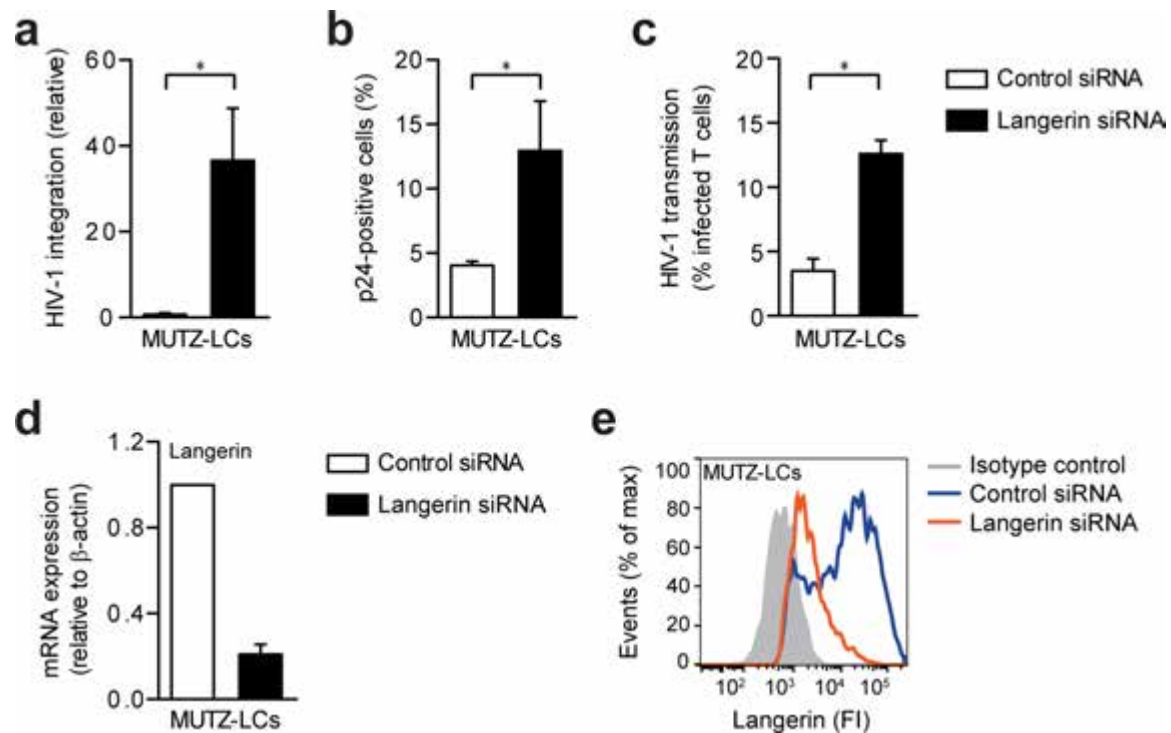
Immunoprecipitation and immunoblotting. Whole-cell extracts were prepared using RIPA lysis buffer supplemented with protease inhibitors. Atg16L1, DC-SIGN, Langerin, p62 and TRIM5 α were immunoprecipitated from 40 μ g of extract with anti- Atg16L1 (PM040; MBL International), DC-SIGN (AZN-D1)¹⁹, Langerin (10E2)⁵, p62 (ab56416; Abcam), TRIM5 α (ab109709; Abcam), mouse IgG1 isotype control (MOPC-21; BD Pharmingen), mouse IgG2a isotype control (IC003A; R&D systems) and rabbit IgG control (sc-2077; Santa Cruz) coated on protein A/G PLUS agarose beads (sc-2003; Santa Cruz), washed twice with ice-cold RIPA lysis buffer and resuspended in Laemmli sample buffer (161-0747, Bio-Rad). Immunoprecipitated samples were resolved by SDS-PAGE (12.5%), and detected by immunoblotting with Atg5 (PM050; MBL), Atg16L1 (MBL), DC-SIGN (551186; BD Biosciences), Langerin (AF2088; R&D Systems), LSP-1 (3812S; Cell Signalling),

TRIM5 α (Abcam) and HIV-p24 (KC57-RD1-PE; Beckman Coulter) antibodies, followed by incubation with Clean-Blot IP Detection Kit-HRP (21232; Thermo Scientific) and ECL detection (34075; Thermo Scientific). Data acquisition was carried out with ImageQuant LAS 4000 (GE Healthcare). Immunoprecipitation with TRIM5 α , Langerin, DC-SIGN, Atg16L1 and p62 pulls-down mostly the TRIM5 α (approximately 56 kDa) form. Relative intensity of the bands was quantified using Image Studio Lite 5.2 software by normalizing β -actin and set at 1 in untreated cells. For gel source data, see Supplementary Fig. 1.

Statistical analysis. Two-tailed Student's *t*-test for paired observations (differences of stimulations within the same donor or cell-type) or unpaired observation (differences between U87 transfectants). Statistical analyses were performed using GraphPad 6.0 software and significance was set at $P < 0.05$ (* $P < 0.05$; ** $P < 0.01$).

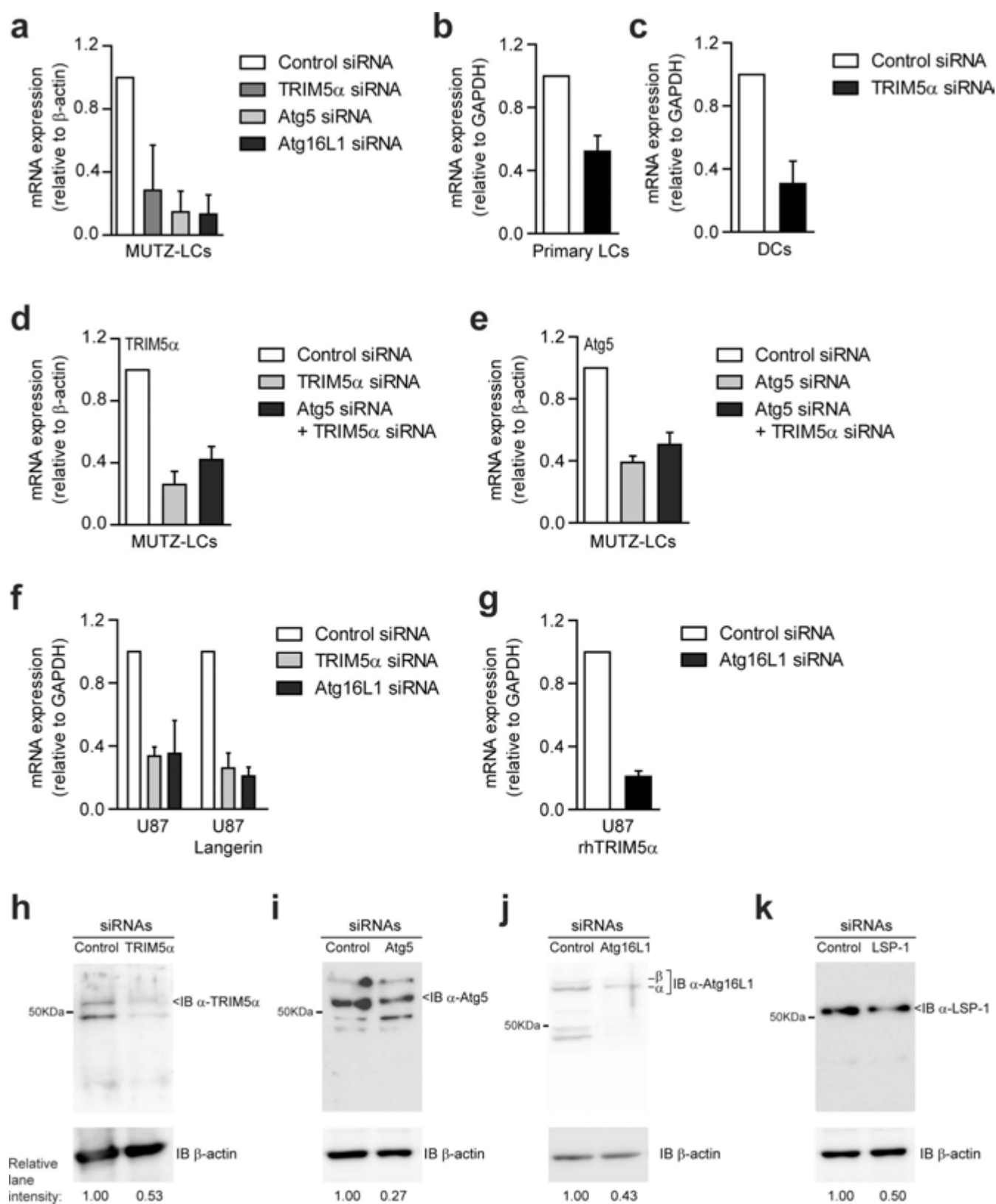
Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

31. Arrighi, J. F. *et al.* DC-SIGN-mediated infectious synapse formation enhances X4 HIV-1 transmission from dendritic cells to T cells. *J. Exp. Med.* **200**, 1279–1288 (2004).
32. Björndal, A. *et al.* Coreceptor usage of primary human immunodeficiency virus type 1 isolates varies according to biological phenotype. *J. Virol.* **71**, 7478–7487 (1997).
33. Setiawan, L. C. & Kootstra, N. A. Adaptation of HIV-1 to rhTrim5 α -mediated restriction *in vitro*. *Virology* **486**, 239–247 (2015).
34. Eng, K. E., Panas, M. D., Karlsson Hedestam, G. B. & McInerney, G. M. A novel quantitative flow cytometry-based assay for autophagy. *Autophagy* **6**, 634–641 (2010).
35. Klionsky, D. J. *et al.* Guidelines for the use and interpretation of assays for monitoring autophagy. *Autophagy* **8**, 445–544 (2012).
36. Chan, L. L. *et al.* A novel image-based cytometry method for autophagy detection in living cells. *Autophagy* **8**, 1371–1382 (2012).



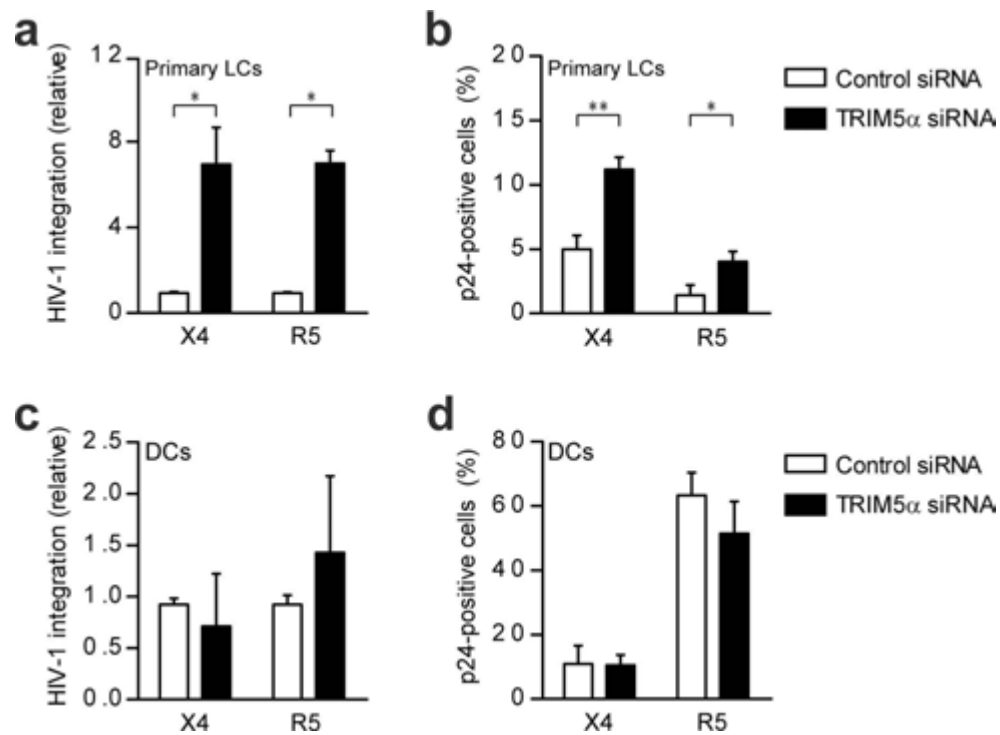
Extended Data Figure 1 | Langerin in MUTZ-LCs restricts HIV-1 integration, infection and transmission to CD4⁺ T cells. **a, b,** HIV-1_{NL4.3} integration (**a**) and infection (**b**) of MUTZ-LCs after Langerin silencing, determined by Alu-PCR (**a**) and intracellular p24 staining (**b**). **c,** HIV-1_{NL4.3-Bal} transmission by MUTZ-LCs after Langerin silencing, determined in LC and T-cell coculture by intracellular p24 staining.

d, e, Silencing was confirmed by real-time PCR (**d**) or by flow cytometer (**e**; representative of $n = 3$). mRNA expression was normalized to β -actin (**d**) and set at 1 in control-siRNA treated cells. * $P < 0.05$ (two-tailed t -test). Data are mean \pm s.d. of three (**a, c, d**) and four (**b**) independent experiments.



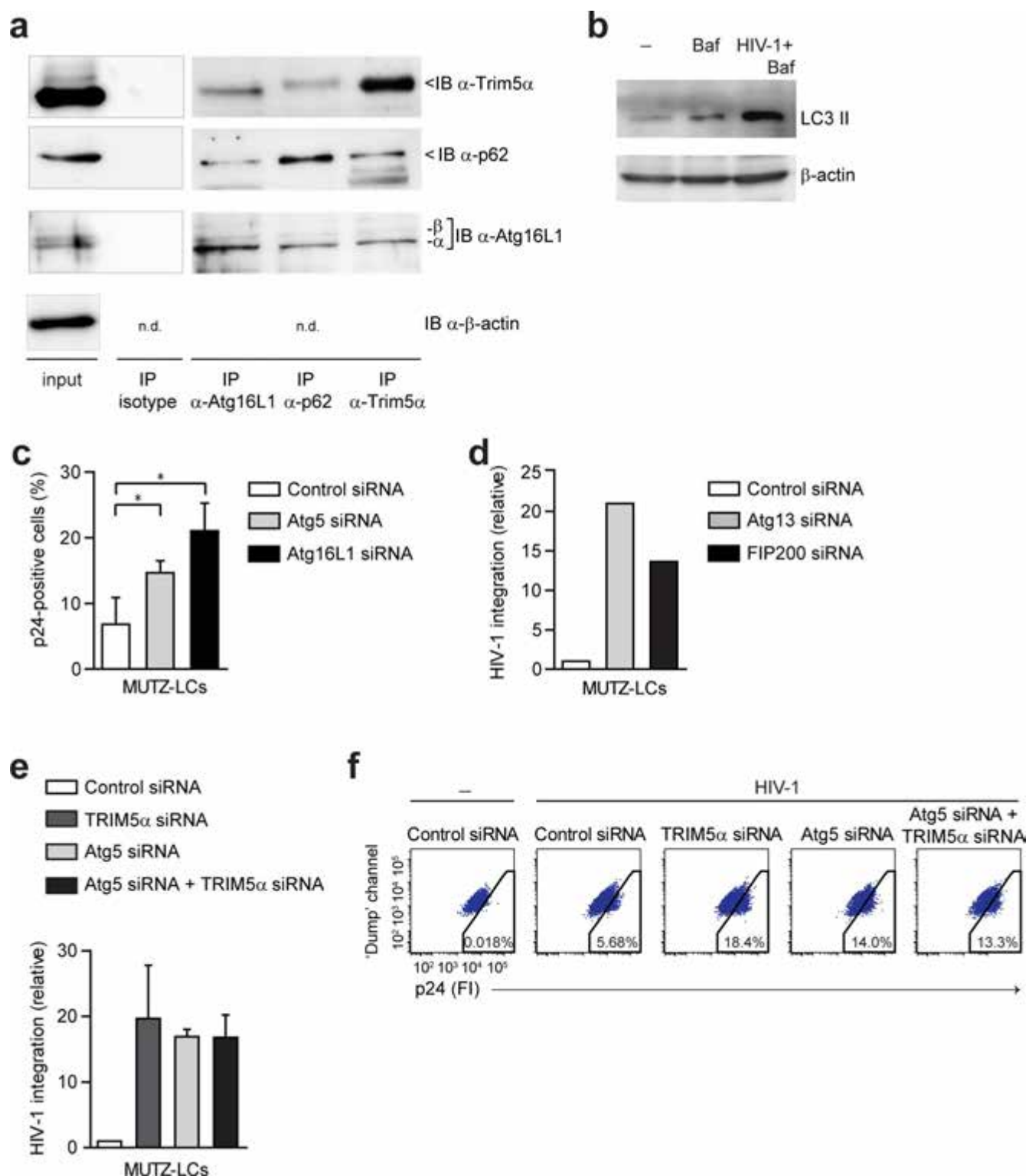
Extended Data Figure 2 | Silencing of TRIM5 α , Atg5, Atg16L1 and LSP-1 by RNA interference. a–k, Indicated proteins were silenced using specific SMARTpools and non-targeting siRNA as a control. Silencing was confirmed by real-time PCR (a–g) or by immunoblotting (β -actin served as loading control; h–k) in MUTZ-LCs (a, d, e, h, i, j), primary LCs (b), DCs (c), CD4⁺CCR5⁺ U87 parental cells (f) or CD4⁺CCR5⁺ U87 cells transduced with either Langerin (f, k) or rhesus TRIM5 α (g). mRNA

expression was normalized to β -actin (a, d, e) or GAPDH (b, c, f, g) and set at 1 in cells treated with control siRNA. Relative abundance of indicated proteins was quantified by normalizing to β -actin and set at 1 in control siRNA treated cells. Representative of $n = 2$ (d, e, h–k). For gel source data, see Supplementary Fig. 1. Data are mean \pm s.d. of three (a, c, f, g) and six (b) independent experiments.



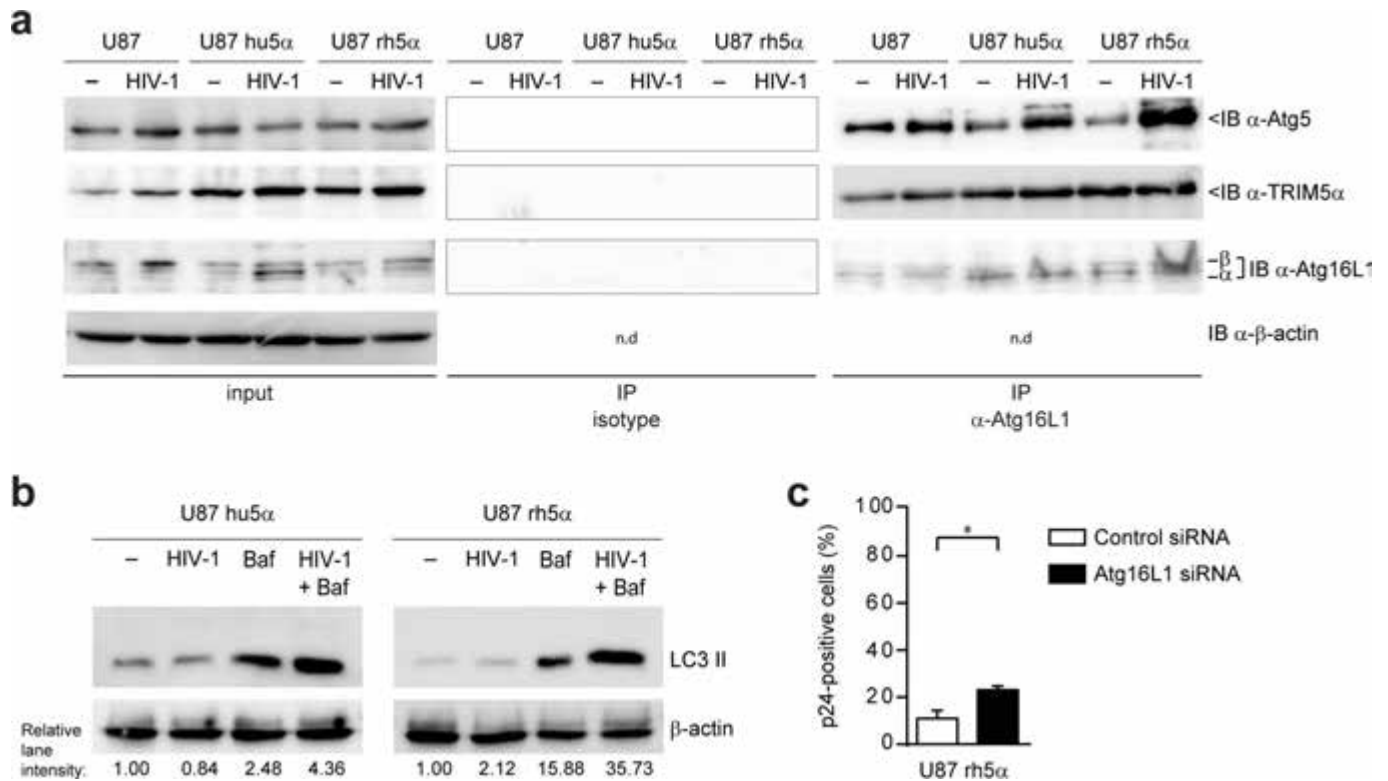
Extended Data Figure 3 | Human TRIM5 α -mediated restriction in LCs or, the lack thereof in DCs, is independent of virus tropism. a–d, HIV-1_{NL4.3} (X4, CXCR4-tropic virus) or HIV-1_{NL4.3-BaL} (R5, CCR5-tropic virus) integration (a, c) and infection (b, d) of primary LCs (a, b) or DCs (b, d)

after TRIM5 α silencing determined by Alu-PCR (a, c) and intracellular p24 staining (b, d). * $P < 0.05$, ** $P < 0.01$ (two-tailed t -test). Data are mean \pm s.d. of three (a–d) independent experiments.



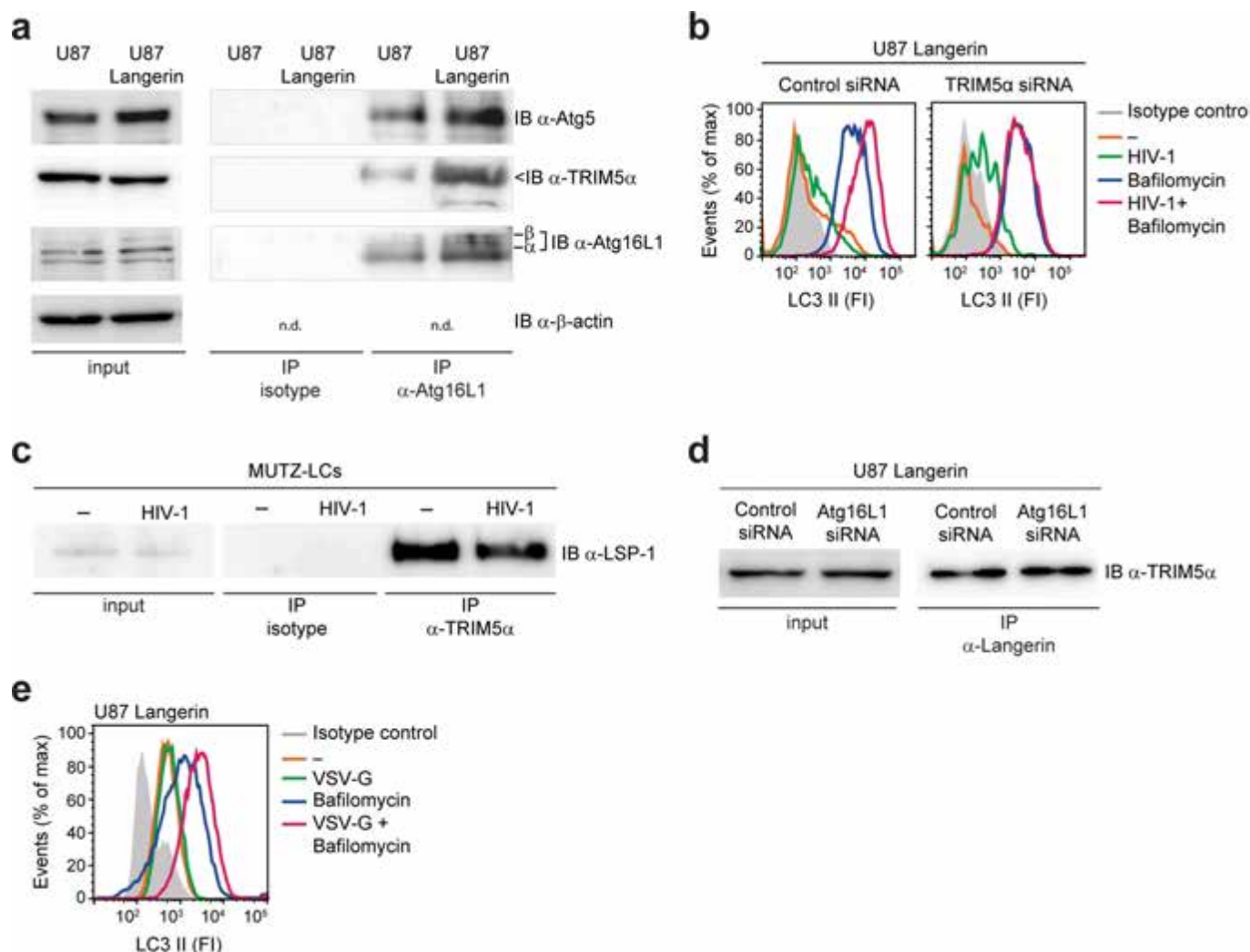
Extended Data Figure 4 | ULK1 complex-dependent autophagy restricts HIV-1 integration in LCs and human TRIM5 α restriction is dependent on Atg5 function. **a**, TRIM5 α , p62 and Atg16L1 in whole-cell lysates of uninfected MUTZ-LCs before (input) or after immunoprecipitation with Atg16L1, p62, TRIM5 α , rabbit IgG control (as control for Atg16L1 and TRIM5 α IP) or mouse IgG2a isotype control (as control for p62 immunoprecipitation), determined by immunoblotting (n.d., not determined). **b**, Autophagy induction in primary LCs pre-treated with bafilomycin followed by incubation with HIV-1_{NL4.3}, determined by

immunoblotting for LC3. For gel source data, see Supplementary Fig. 1. **c**, HIV-1_{NL4.3} infection of MUTZ-LCs after Atg5 or Atg16L1 silencing, determined by intracellular p24 staining. **d**, HIV-1_{NL4.3} integration into MUTZ-LCs after Atg13 or FIP200 silencing, determined by Alu-PCR. **e**, **f**, HIV-1_{NL4.3} integration (e) or infection (f) of MUTZ-LCs after Atg5, TRIM5 α silencing or simultaneously with Atg5 and TRIM5 α silencing, determined by Alu-PCR (e) and intracellular p24 staining (f). Data are representative of three (a) or two (b, d–f) experiments and mean \pm s.d. of four independent experiments (c).



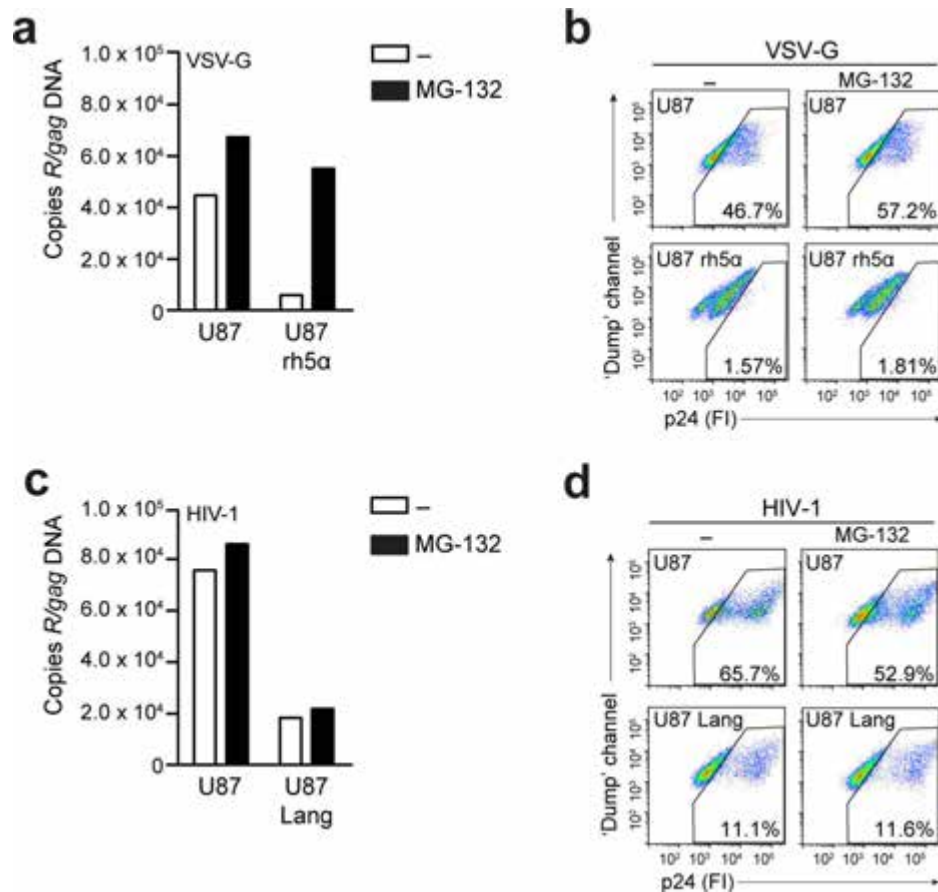
Extended Data Figure 5 | Increased Atg5 recruitment into TRIM5α–Atg16L1 complex scaffold in CD4⁺CCR5⁺ U87 transfectants. a, Atg5, TRIM5α and Atg16L1 in whole-cell lysates of CD4⁺CCR5⁺ U87 parental cells (U87) or transduced with either human TRIM5α (U87 hu5α) or rhesus TRIM5α (U87 rh5α) infected with HIV-1_{NL4.3-BaL} before (input) or after immunoprecipitation with Atg16L1 or rabbit IgG control, determined by immunoblotting (n.d., not determined). **b**, Autophagy induction in U87 transfectants with bafilomycin followed by incubation with HIV-1_{SF162},

determined by immunoblotting for LC3 (autophagy induction in control CD4⁺CCR5⁺ U87 parental cells presented in Fig. 3o). Relative abundance of LC3 II determined by normalizing to β-actin. Representative of $n = 2$ (**a**, **b**). For gel source data, see Supplementary Fig. 1. **c**, HIV-1_{SF162} infection of CD4⁺CD5⁺ U87 cells transduced with rhesus TRIM5α after Atg16L1 silencing, determined by intracellular p24 staining. * $P < 0.05$ (t -test). Data are mean \pm s.d. of three (**c**) independent experiments.



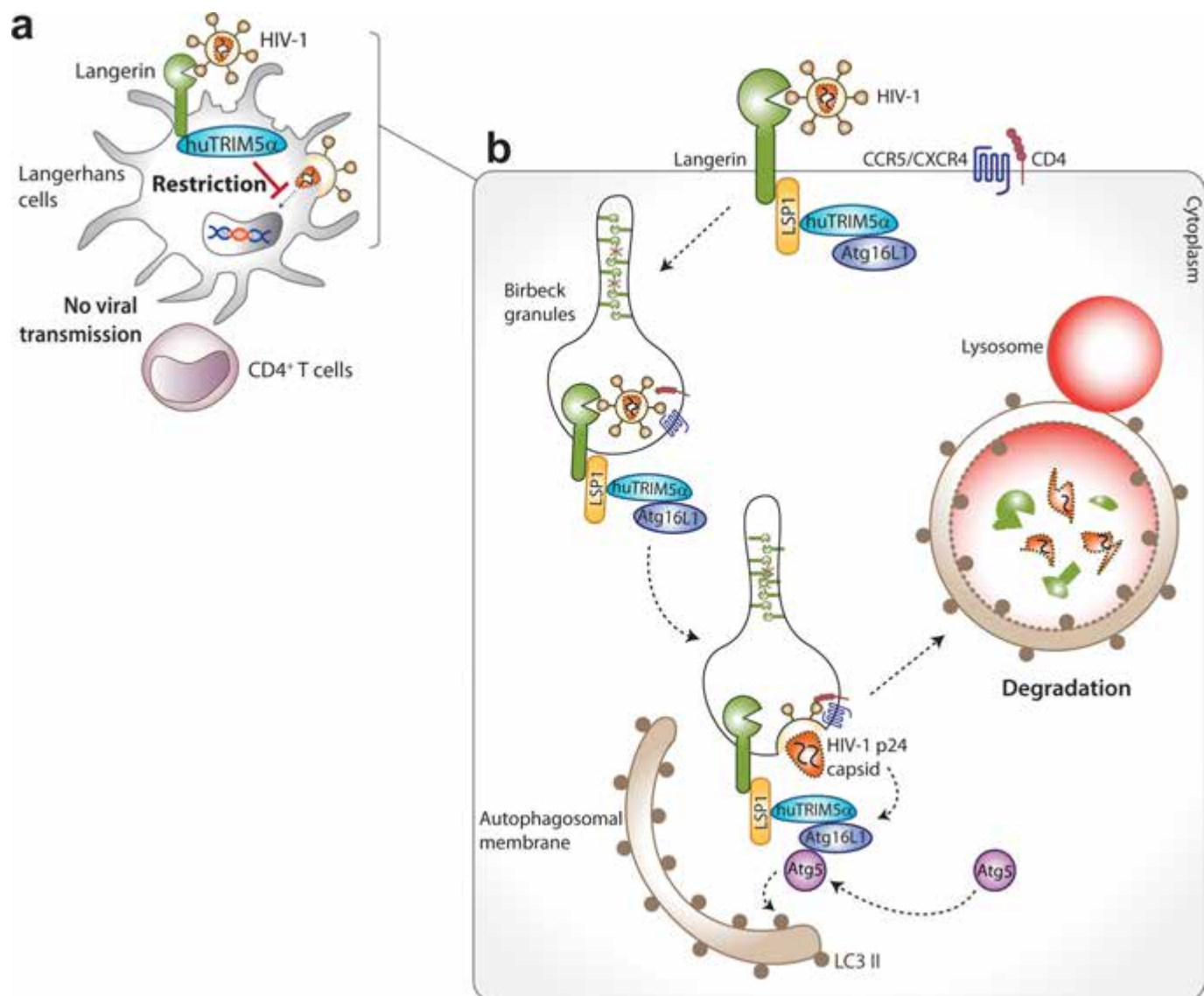
Extended Data Figure 6 | Human TRIM5 α induces autophagy upon HIV-1 exposure in Langerin⁺ U87 transfectant and interacts with Langerin through LSP-1, but not Atg16L1. **a**, Atg5, TRIM5 α and Atg16L1 in whole-cell lysates of CD4⁺CCR5⁺ U87 parental cells (U87) or transduced with Langerin (U87 Langerin) before (input) or after immunoprecipitation with Atg16L1 or rabbit IgG control, determined by immunoblotting (n.d., not determined). **b**, Autophagy levels in Langerin⁺ U87 transfectant after TRIM5 α silencing, pre-treated with bafilomycin followed by incubation with HIV-1_{NL4.3-BaL}, determined by intracellular

LC3 II levels by flow cytometer. **c**, LSP-1 in whole-cell lysates of MUTZ-LCs infected with HIV-1_{NL4.3} before (input) or after immunoprecipitation with TRIM5 α or rabbit IgG control. **d**, TRIM5 α in whole-cell lysates of Langerin⁺ U87 transfectant after Atg16L1 silencing before (input) or after immunoprecipitation with Langerin, determined by immunoblotting. For gel source data, see Supplementary Fig. 1. **e**, Autophagy induction in Langerin⁺ U87 transfectant pre-treated with bafilomycin followed by incubation with VSV-G-pseudotyped HIV-1, determined by intracellular LC3 II levels. Data are representative of two experiments (**a**–**e**).



Extended Data Figure 7 | Proteasome inhibition does not relieve Langerin-mediated restriction of HIV-1 reverse-transcription products nor infection. a–d, *R/gag* proviral DNA levels (a, c) and HIV-1 infection (b, d) in CD4⁺CCR5⁺ U87 parental cells (U87) or cells transduced with either rhesus TRIM5α (U87 rh5α) or Langerin (U87 Lang) after

pre-treatment with proteasome inhibitor MG-132 and infected with VSV-G-pseudotyped HIV-1 (a, b; VSV-G) or HIV-1_{NL4.3-BaL} (c, d; HIV-1), determined by qPCR (a, c) and intracellular p24 staining (b, d). Data are representative of two experiments (a–d).



Extended Data Figure 8 | Langerin-controlled human TRIM5 α restriction mechanism in Langerhans cells. a, HIV-1 binding to Langerin in Langerhans cells drives human TRIM5 α -mediated restriction of viral integration, HIV-1 infection and HIV-1 transmission to CD4⁺ T cells. **b**, Langerin associates at steady-state with LSP-1–TRIM5 α –Atg16L1 complex. Capture of HIV-1 by Langerin targets internalization of the

incoming virus into Birbeck granules. Upon viral fusion, human TRIM5 α mediates recruitment of Atg5 to TRIM5 α –Atg16L1–HIV-1p24 capsid complex, which promotes lipidation of LC3 (LC3 II) and thereby elicit autophagosome formation. Vesicles containing Langerin–HIV-1 capsid complexes are subsequently targeted into autophagosomes for lysosomal degradation, which prevents infection of Langerhans cells.

Extended Data Table 1 | Primer sequences used for mRNA expression and HIV-1 integration assay

mRNA expression primer sequences		
Gene product	Forward primer	Reverse primer
Atg5	TCATTCAGAAGCTGTTTCGTCC	CCCCATCTTCAGGATCAATAGC
Atg16L1	TGCTCCCGTGATGACTTGC	CAACTCTGGTCCAGTCAGAGCC
TRIM5 α	AGAACATACGGCCTAATCGGC	CAACTTGACCTCCCTGAGCTTC
GAPDH	CCATGTTTCGTCATGGGTGTG	GGTGCTAAGCAGTTGGTGGTG
β -actin	GCTCCTCCTGAGCGCAAG	CATCTGCTGGAAGGTGGACA

HIV-1 integration primer sequences	
Primer	Primer sequence
HIV-1 LTR R ⁺ (forward)	<u>ATGCCACGTAAGCGAAACTG</u> CTGGCTAACTAGGGAACCCACTG
Alu (reverse)	TCCCAGCTACTGGGGAGGCTGAGG
Second-round marker (forward)	ATGCCACGTAAGCGAAACTG
HIV-1 LTR U5 (reverse)	CACACTGACTAAAAGGGTCTGAGG
GAPDH DNA (forward)	TACTAGCGGTTTTACGGGCG
GAPDH DNA (reverse)	TCGAACAGGAGGAGCAGAGAGCGA

*Marker sequence at 5' end underlined.

Structure of photosystem II and substrate binding at room temperature

Iris D. Young^{1*}, Mohamed Ibrahim^{2*}, Ruchira Chatterjee^{1*}, Sheraz Gul¹, Franklin D. Fuller¹, Sergey Koroidov³, Aaron S. Brewster¹, Rosalie Tran¹, Roberto Alonso-Mori⁴, Thomas Kroll^{5,6}, Tara Michels-Clark¹, Hartawan Laksmono⁵, Raymond G. Sierra^{4,5}, Claudiu A. Stan⁵, Rana Hussein², Miao Zhang², Lacey Douthit¹, Markus Kubin⁷, Casper de Lichtenberg³, Long Vo Pham³, Håkan Nilsson³, Mun Hon Cheah³, Dmitriy Shevela³, Claudio Saracini¹, Mackenzie A. Bean¹, Ina Seuffert², Dimosthenis Sokaras⁶, Tsu-Chien Weng^{6,†}, Ernest Pastor¹, Clemens Weninger⁵, Thomas Fransson⁵, Louise Lassalle¹, Philipp Bräuer^{8,9}, Pierre Aller⁹, Peter T. Docker⁹, Babak Andi¹⁰, Allen M. Orville⁹, James M. Glowinski⁴, Silke Nelson⁴, Marcin Sikorski⁴, Diling Zhu⁴, Mark S. Hunter⁴, Thomas J. Lane⁴, Andy Aquila⁴, Jason E. Koglin⁴, Joseph Robinson⁴, Mengning Liang⁴, Sébastien Boutet⁴, Artem Y. Lyubimov^{11,12}, Monarin Uervirojnangkoorn^{11,12}, Nigel W. Moriarty¹, Dorothee Liebschner¹, Pavel V. Afonine¹, David G. Waterman^{13,14}, Gwyndaf Evans⁹, Philippe Wernet⁷, Holger Dobbek², William I. Weiss^{11,15,16}, Axel T. Brunger^{11,12,15,16}, Petrus H. Zwart¹, Paul D. Adams^{1,17}, Athina Zouni², Johannes Messinger^{3,18}, Uwe Bergmann⁵, Nicholas K. Sauter¹, Jan Kern^{1,4}, Vittal K. Yachandra¹ & Junko Yano¹

Light-induced oxidation of water by photosystem II (PS II) in plants, algae and cyanobacteria has generated most of the dioxygen in the atmosphere. PS II, a membrane-bound multi-subunit pigment protein complex, couples the one-electron photochemistry at the reaction centre with the four-electron redox chemistry of water oxidation at the Mn₄CaO₅ cluster in the oxygen-evolving complex (OEC). Under illumination, the OEC cycles through five intermediate S-states (S₀ to S₄)¹, in which S₁ is the dark-stable state and S₃ is the last semi-stable state before O–O bond formation and O₂ evolution^{2,3}. A detailed understanding of the O–O bond formation mechanism remains a challenge, and will require elucidation of both the structures of the OEC in the different S-states and the binding of the two substrate waters to the catalytic site^{4–6}. Here we report the use of femtosecond pulses from an X-ray free electron laser (XFEL) to obtain damage-free, room temperature structures of dark-adapted (S₁), two-flash illuminated (2F; S₃-enriched), and ammonia-bound two-flash illuminated (2F-NH₃; S₃-enriched) PS II. Although the recent 1.95 Å resolution structure of PS II at cryogenic temperature using an XFEL⁷ provided a damage-free view of the S₁ state, measurements at room temperature are required to study the structural landscape of proteins under functional conditions^{8,9}, and also for *in situ* advancement of the S-states. To investigate the water-binding site(s), ammonia, a water analogue, has been used as a marker, as it binds to the Mn₄CaO₅ cluster in the S₂ and S₃ states¹⁰. Since the ammonia-bound OEC is active, the ammonia-binding Mn site is not a substrate water site^{10–13}. This approach, together with a comparison of the native dark and 2F states, is used to discriminate between proposed O–O bond formation mechanisms.

Diffraction to 2.0 Å was observed at room temperature and structural datasets for the S₁ and 2F states of PS II (Fig. 1a, b and Extended Data Fig. 1a) under different buffer conditions were obtained at 3.0 to 2.25 Å resolution (see Methods, Extended Data Table 1 and Supplementary Tables 1–4). The packing of the PS II dimers (Fig. 1c) and unit cell

dimensions differ significantly from those reported in ref. 7 (Extended Data Fig. 1). For the illuminated data, the unit cell dimensions remained the same as for the S₁-state data, in contrast to a recent report¹⁴ (see also ref. 15). Examples of the electron densities ($2mF_o - DF_c$ and polder bulk solvent-excluding omit maps, see Methods) for several structural groups are shown in Fig. 1d–f and Extended Data Fig. 2 for the S₁ state (3.0 Å resolution) and the 2F state, with and without ammonia (at 2.8 and 2.25 Å resolution, respectively).

Compared to the cryogenic S₁ XFEL structure⁷, large-scale rotation of the monomers relative to each other by approximately 0.6° in the dimeric complex is observed in our room temperature S₁ XFEL structure. Within each monomer, the locations and orientations of transmembrane helices (TMHs) also differ, resulting in expansion of each monomer within the membrane (Fig. 2a and Extended Data Fig. 3). We also observe systematic elongation of the centre-to-centre distances between the cofactors of the electron transport chain by 0.1–0.4 Å (Fig. 2b and Extended Data Fig. 4a) and expansion of distances between adjacent chlorophyll centres involved in energy transfer in the antennae (CP43, CP47) (Fig. 2c, Extended Data Fig. 4b–d and Supplementary Table 5) by 1–3.5% (0.3–0.9 Å) compared to the cryogenic S₁ XFEL structure⁷. Changes of this magnitude are larger than the errors for the cofactors in our structure (Methods and Supplementary Table 5). Comparison of our room temperature S₁ XFEL structure with the cryogenic S₁ XFEL structure⁷ and a cryogenic S₁ synchrotron radiation structure¹⁶ that featured dimer packing similar to our room temperature XFEL structure (Extended Data Figs 3, 4 and Supplementary Information) reveals that the majority of changes in TMH positions and cofactor distances are the result of the difference in data collection temperatures. Differences in the equilibrium distances have a profound influence on the calculated electron and excitation transfer rates in PS II. For example, extension of the pheophytin (Pheo)_{D1}–Q_A (primary acceptor plastoquinone bound to PS II) distance by 0.2 Å leads to a reduction in the calculated electron transfer

¹Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ²Institut für Biologie, Humboldt-Universität zu Berlin, D-10099 Berlin, Germany. ³Institutionen för Kemi, Kemiskt Biologiskt Centrum, Umeå Universitet, SE 90187 Umeå, Sweden. ⁴LCLS, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. ⁵Stanford PULSE Institute, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. ⁶SSRL, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. ⁷Institute for Methods and Instrumentation on Synchrotron Radiation Research, Helmholtz Zentrum, 14109 Berlin, Germany. ⁸Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK. ⁹Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0DE, UK. ¹⁰National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, New York 11973, USA. ¹¹Department of Molecular and Cellular Physiology, Stanford University, Stanford, California 94305, USA. ¹²Howard Hughes Medical Institute, Stanford University, California 94305, USA. ¹³STFC Rutherford Appleton Laboratory, Didcot, Oxfordshire OX11 0QX, UK. ¹⁴CCP4, Research Complex at Harwell, Rutherford Appleton Laboratory, Didcot, Oxfordshire OX11 0FA, UK. ¹⁵Department of Photon Science, Stanford University, Stanford, California 94305, USA. ¹⁶Department of Structural Biology, Stanford University, Stanford, California 94305, USA. ¹⁷Department of Bioengineering, University of California Berkeley, Berkeley, California 94720, USA. ¹⁸Department of Chemistry, Molecular Biomimetics, Ångström Laboratory, Uppsala University, SE 75237 Uppsala, Sweden. [†]Present address: Center for High Pressure Science & Technology Advanced Research, Pudong, Shanghai 201203, China.

*These authors contributed equally to this work.

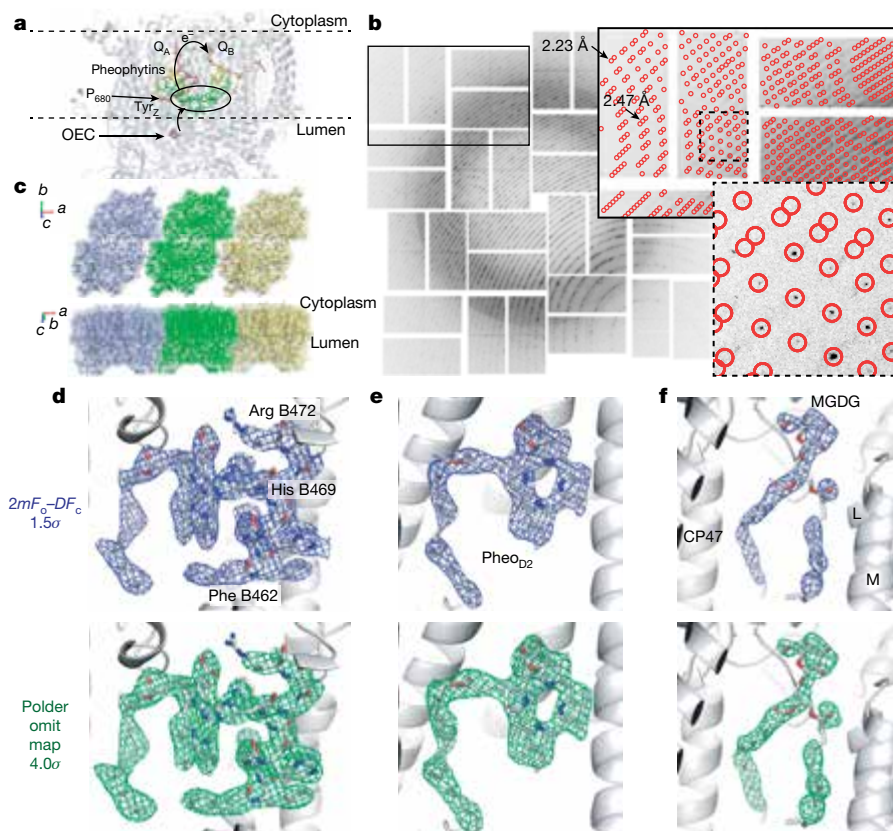


Figure 1 | Crystal structure of PS II in the dark (S_1) state collected at room temperature with femtosecond XFEL pulses. **a**, Schematic overview of the reaction centre of PS II showing cofactors involved in light-induced electron transfer and water oxidation. **b**, Diffraction pattern of a PS II microcrystal. Insets show enlarged views with the indexing solution. **c**, Arrangement of dimers in the unit cells of the crystals. Three

dimers (blue, green, yellow) with the view onto the membrane plane from the luminal side (top) and with the view along the membrane plane (bottom). **d–f**, Examples of omit maps with the model (grey). **d**, TMH 6 of CP47 with a Chl. **e**, Pheophytin in subunit D2. **f**, Monogalactosyl diacylglycerol (MGDG) lipid next to CP47.

rates¹⁷ of about 25% and elongation of the distance between connecting chlorophylls of the antennae and reaction centre pigments from 20.5 to 20.9 Å yields a reduction in excitation transfer rates¹⁸ of about 10%.

The room temperature structure exhibits an increase of about 50% in altered rotamers of side-chains relative to the differences between the cryogenic XFEL and synchrotron radiation structures (Fig. 2d, e), indicating that the rotamer distribution is affected by temperature, as observed for other proteins⁸. These differences are observed nearly exclusively in solvent-exposed regions of the complex (Fig. 2d), and they can be explained by (a) high flexibility of the residues at room temperature leading to no or only weak electron density; (b) the presence of multiple conformers at room temperature (for example, Arg B476 in Fig. 2e); or (c) stabilization of a different conformer at room temperature (for example, Phe B479 in Fig. 2e).

In our room temperature S_1 and 2F-NH₃ XFEL structures (3.0 and 2.8 Å resolution, respectively), around 120 waters per dimer were resolved, with omit maps indicating the presence of additional waters (Methods, Extended Data Fig. 5 and Supplementary Table 6). In the higher-resolution native 2F room temperature XFEL structure, ~1,200 waters were observed. The observation of key water molecules linking the OEC to possible proton channels¹⁹ in our data (Extended Data Fig. 5) indicates that the channels postulated based on cryogenic structures^{7,19} may be relevant in PS II under physiological conditions.

Electron density maps for the OEC are shown in Fig. 3a, b and Extended Data Fig. 6 for the S_1 (3.0 Å) and 2F data (2.25 Å). To minimize model bias, metal–metal distances in the OEC (Fig. 3c) were calculated from individual metal omit maps. The similarity between the distances in the S_1 room temperature structure and those in the cryogenic S_1 XFEL structure⁷ (see error estimates in Methods) confirms that the observed metal arrangement at 100 K⁷ is valid for the

physiological S_1 state at room temperature. Although it is difficult to determine bridging oxygen positions in multimetallic clusters such as the OEC, because the electron density of the oxygens tends to be overwhelmed by the nearby metal density, oxygen positions modelled in the room temperature S_1 state match well with those of the cryogenic S_1 XFEL structure. On the basis of the Mn4–O5 (~2.3 Å) and Mn1–O5 (~2.7 Å) distances, however, we concluded, as suggested by X-ray spectroscopy³, electron paramagnetic resonance (EPR)²⁰ and theory²¹, that O5 (Fig. 3a) is bound to the Mn4 site, but not bound to Mn1 in the S_1 state. This leads to an open cubane structure for the Mn₄CaO₅ unit.

In situ illumination of dark-adapted crystals at room temperature led to the formation of 2F crystals in which the S_3 state was predominant (Methods and Extended Data Fig. 7). In the 2F room temperature XFEL data, the cluster maintains its overall structure (Fig. 3b, c and Extended Data Fig. 6b), but changes are observed in some of the atomic positions, such as those of O4 and Mn4. The Mn3–Mn4 distance is shortened by about 0.1 Å and there is a twist of the Mn3–O4–Mn4–O5 plane in relation to the Mn1–Mn2–Mn3 core structure (Fig. 3d). While a distance uncertainty of 0.10–0.15 Å remains at 2.25 Å resolution, the trend in metal–metal distance changes does not support S_3 models based on closed cube geometries²² (for example, Extended Data Fig. 8a,d) that require more than 0.5 Å elongation of Mn3–Mn4 or compression of Mn1–Mn3 distances. Similarly, our data do not support models in which a new water or hydroxo binds to Mn1^{20,21} (Extended Data Fig. 8b), as no corresponding electron density or distance changes were observed (Fig. 3e, f and Extended Data Fig. 6b).

All four water ligands to the OEC (W1–W4, Extended Data Fig. 1a) are clearly visible in our S_3 -enriched 2F room temperature XFEL data, within 0.2–0.4 Å of their positions in the cryogenic S_1 XFEL structure; for example, W3 is displaced towards Glu189 and W4 towards Asp170

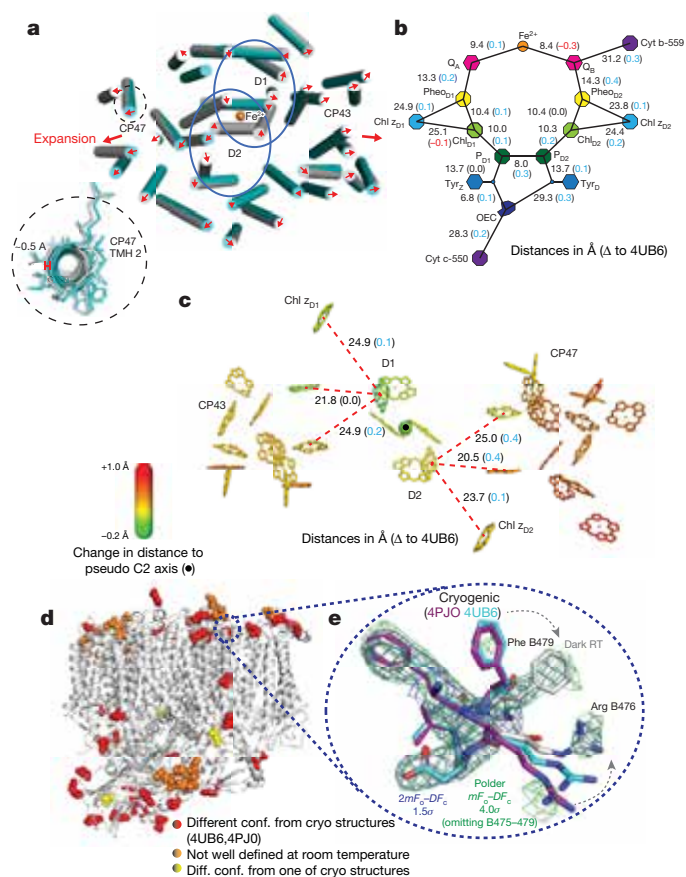


Figure 2 | Comparison of the room temperature and cryogenic structures in the dark state. **a**, Location of TMH in one monomer at room temperature (grey cylinders) and the cryogenic structure⁷ (cyan). View is onto the membrane plane from the cytoplasmic side. One TMH is shown enlarged in the inset to illustrate the shift between the cryogenic and room temperature structures. **b**, Distances between the cofactors of the electron transport chain in the dark structure, with differences from the cryogenic structure⁷ in parentheses. **c**, Changes in Chl-Phen distances, represented by colour. **d**, Location of residues that show different side chain orientations at room temperature compared to the cryogenic structures^{7,16}. **e**, Examples of different side chain positions in the room temperature structure (grey) with the cryogenic structures (cyan⁷, purple¹⁶). Dashed lines indicate the location of these residues in the PS II complex.

(Fig. 3b, d and Extended Data Fig. 5e; note that the W3 position in our room temperature S₁ structure is shifted by about 0.8 Å from that of the cryogenic S₁ XFEL structure, Extended Data Fig. 5d). No additional water or hydroxo ligand to the OEC was observed in the 2F data. Comparison of the roughly 20 waters within 7 Å of the OEC indicated that three of the second-sphere waters from the cryogenic S₁ XFEL structure (A571, A588 and C665) could not be located in the 2F structure, probably owing to their mobility at room temperature or changes in the water arrangement around the OEC upon formation of the S₃ state (Extended Data Fig. 5e).

There are no large movements of the side chains surrounding the OEC—for example, Asp 61, Asp 170 and Val 185 of D1—between the dark and native 2F structures (Fig. 3a, b, d) of the sort predicted in previous studies^{14,23}. Instead, small movements on the order of 0.3 Å are noticeable in several side chains, in agreement with the data from ref. 24.

At the electron-acceptor site, no large differences are observed for the mobile quinone Q_B between the dark and native 2F structures. This is in line with the expectation that after two flashes Q_BH₂ is released and the Q_B pocket is filled with a new plastoquinone within the 500 ms period before sampling with the XFEL pulse (Extended Data Fig. 7a, b).

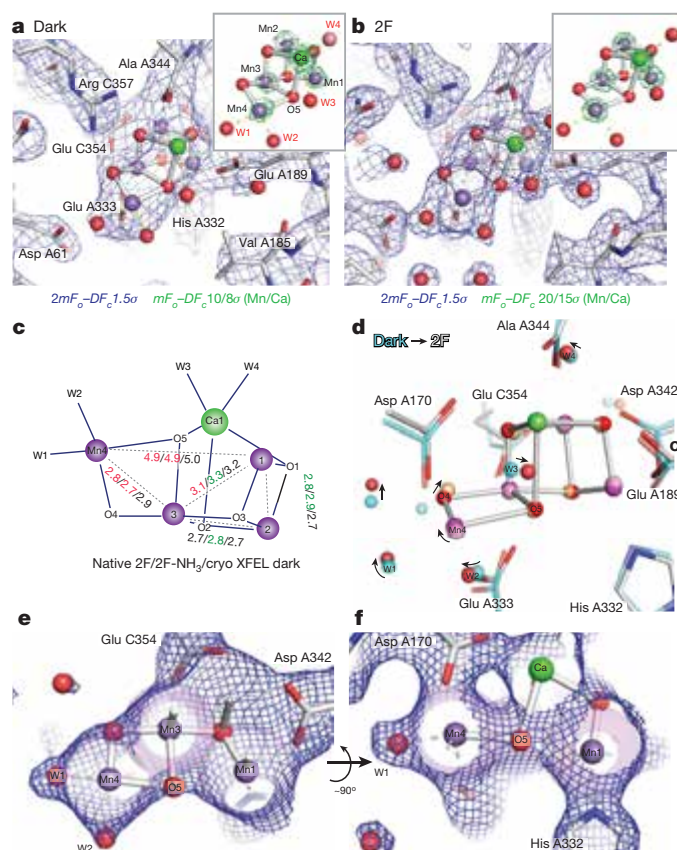


Figure 3 | The oxygen evolving complex. **a**, **b**, Electron density around the OEC of the dark (a) and 2F (b) room temperature structures. Insets, individual metal atom omit maps. **c**, Mn-Mn distances obtained from metal omit maps for the 2F and 2F-NH₃ data compared with the distances from ref. 7. **d**, Perturbations in the vicinity of the Mn₄CaO₅ cluster in the 2F state (grey) in comparison to ref. 7 (cyan). **e**, **f**, Electron density in the Mn1-O5-Mn4 region in two different orientations. Density (blue mesh, half the grid spacing compared to other maps shown) is contoured at 1.8σ, matching the van der Waals radius (light magenta spheres) of Mn. Clear density protruding beyond the van der Waals volume of the Mn is visible for metal bound waters (for example, W1, W2) but no extra density is seen around Mn1, a proposed location of an inserted water in the Mn₄CaO₅ cluster in the S₃ state.

Insight into which of the water or hydroxo ligands of the Mn₄CaO₅ cluster are the substrate waters can be obtained by analysing differences induced by ammonia binding (see Supplementary Information). Ammonia is known to bind to the Mn₄CaO₅ cluster at a non-substrate water binding site only upon illumination¹⁰. We therefore obtained 2F room temperature XFEL data at pH 7.5 (2F-NH₃, 2.8 Å resolution) in the presence of 100 mM (NH₄)₂SO₄, and compared it with the native 2F data. In the 2F-NH₃ structure, about 75% of PS II centres are in either the S₂ or S₃ state (Methods and Extended Data Fig. 7f–i). Binding of ammonia to Mn was confirmed by the altered S₂ EPR multiline signal²⁵ (Extended Data Fig. 7d, e). Although the Cl[−]-binding site was reported as a second, possibly inhibitory ammonia-binding site (see Supplementary Information), we can exclude substitution of Cl[−] by ammonia in our samples based on oxygen-evolution activity (Extended Data Fig. 7f–i) and inspection of the electron density (Extended Data Fig. 9a, b). As a direct distinction between NH₃ and H₂O cannot be drawn from the data, we examined the structure around possible binding sites in the 2F-NH₃ data in comparison with the native 2F structure.

One of the proposed ammonia-binding sites is the μ-oxo bridging ligand O5 (Fig. 4A). Elongation of the Mn3–Mn4 distance is expected upon replacement of a μ-oxo bridge with an amido or imido-bridge¹¹ or a significant alteration of the core structure²⁶; for example, changing

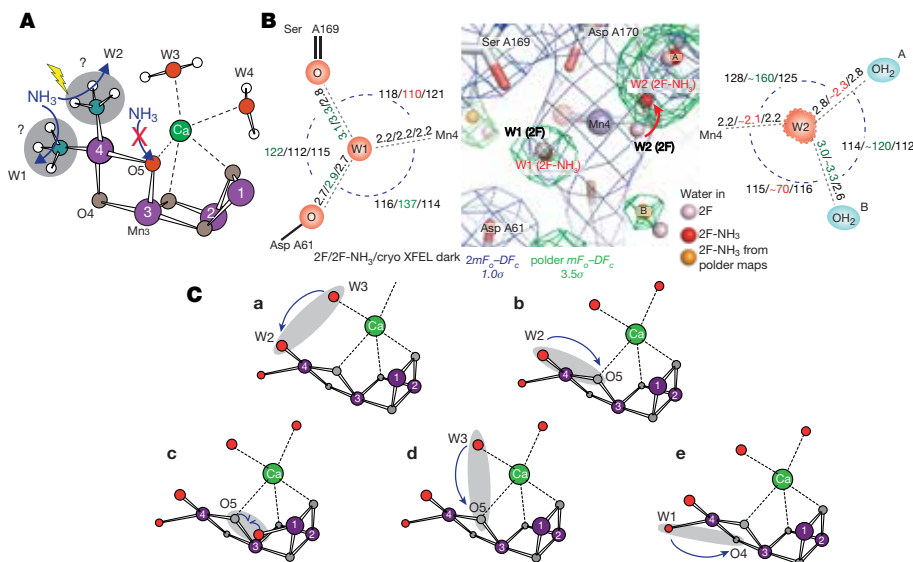


Figure 4 | Ammonia binding sites to the OEC and O–O bond formation mechanisms. **A**, Schematic of the OEC showing suggested locations of ammonia binding upon illumination. **B**, Electron density in the vicinity of Mn4, showing waters W1 and W2 and their surroundings in the 2F and 2F-NH₃ data. Left, Schematic of the bonding geometry of W1 and W2 in ref. 7; right, the 2F and 2F-NH₃ room temperature structures; note that the 2F-NH₃ W2 position is not well defined and angles and distances are

only approximations. **C**, Proposed O–O bond formation mechanisms. **Ca**, Water-nucleophilic attack of Ca-bound W3 onto W2 (for example, Mn(V)-oxo in S₄); **Cb**, coupling between W2 and O5; **Cc**, oxo-oxyl radical coupling mechanism between a hydroxo bound in the S₂ → S₃ transition (hydroxyl in S₄) to Mn1 and the O5-bridge; **Cd**, nucleophilic attack of Ca-bound W3 onto O5; **Ce**, oxo-oxyl coupling between W1 and O4.

from a di- μ -oxo to a mono- μ -oxo interaction upon loss of the Mn3–O5–Mn4 μ -oxo bridge. No such elongation was observed in the 2F-NH₃ data and therefore we eliminate this possibility, in agreement with EPR data¹². Alternatively, ammonia could replace a terminal water ligand on Mn4 (W1 or W2; Fig. 4A), with W1 being favoured based on the interpretation of EPR data^{12,13}. The W1 position in the 2F-NH₃ data is very similar to that in the native 2F structure, while the placement of W2 reveals a shift of its position (Fig. 4B and Extended Data Fig. 9). Of the waters hydrogen-bonded to W2, H₂O^A (Fig. 4B) is in a position similar to that in the native 2F and cryogenic S₁ XFEL structures, whereas H₂O^B is less well defined. A larger tilt of the W3 and W4 axis is also visible in the 2F-NH₃ data as compared with the native 2F data. We note that the native 2F data were obtained by lowering the pH from 7.5 to 6.5; thus, an alternative explanation for the observed change in the W2 site is that W2 is a hydroxide at pH 7.5, while it is fully protonated at pH 6.5.

Modelling the binding of NH₃ in place of W1 in the S₂-state^{12,27} predicted either changes in the Mn3–Mn4 distance by 0.4 Å and displacement of the ammonia nitrogen with respect to the W1 position by about 1 Å²⁷ or ammonia at a position very similar to W1 in the structure shown in ref. 7 with minimal changes in the Mn3–Mn4 distance and a small movement of W2¹². As our data do not show elongation of the Mn3–Mn4 distance, a scenario as postulated in ref. 27 is unlikely. This leaves two options for NH₃-binding in the S₃ state, either at W1 with only minimal changes in the metal positions and ligand environment or at W2 (Fig. 4A).

W2 is less integrated than W1 in a strong hydrogen-bonding network (as also seen in ref. 7), suggesting that there will be easier exchange of W2 if it is bound as fully protonated H₂O. On the other hand, the weaker H-bonding makes ammonia binding at the W2 site difficult to reconcile with the highly anisotropic nuclear quadrupole parameter of the bound ammonia¹³. We note that our data do not exclude ammonia moving from the W1 to W2 site or detaching from Mn during the S₂ → S₃ transition^{27,28}.

Ammonia binding does not significantly affect the substrate water exchange rates in the S₂ and S₃ states^{10,12} (Extended Data Fig. 7). Therefore, if the structural change at W2 is caused, directly or indirectly, by ammonia binding, then W2 is not a likely substrate binding site in

the S₃-state. In the context of the proposed mechanisms^{4,29–31} (Fig. 4C), this would disfavour O–O bond formation via nucleophilic attack from Ca-bound W3 onto W2 (Fig. 4C, a) and other mechanisms that use W2 (for example, Fig. 4C, b). As we did not find evidence for the presence of an additional water or hydroxo near Mn1 in the 2F samples, our data do not support direct coupling between a newly bound water-derived ligand in the S₃ state on Mn1 and O5 (Fig. 4C, c; see also ref. 2). This will leave possible mechanisms such as O–O bond formation between W3 and O5 (Fig. 4C, d), between W1 and O4 (Fig. 4C, e), or other relevant mechanisms.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 June; accepted 14 October 2016.

Published online 21 November 2016.

- Kok, B., Forbush, B. & McGloin, M. Cooperation of charges in photosynthetic O₂ evolution-I. A linear four step mechanism. *Photochem. Photobiol.* **11**, 457–475 (1970).
- Cox, N. & Messinger, J. Reflections on substrate water and dioxygen formation. *Biochim. Biophys. Acta* **1827**, 1020–1030 (2013).
- Yano, J. & Yachandra, V. Mn₄Ca cluster in photosynthesis: where and how water is oxidized to dioxygen. *Chem. Rev.* **114**, 4175–4205 (2014).
- Messinger, J., Badger, M. & Wydrzynski, T. Detection of one slowly exchanging substrate water molecule in the S₃ state of photosystem II. *Proc. Natl Acad. Sci. USA* **92**, 3209–3213 (1995).
- Hillier, W. & Wydrzynski, T. ¹⁸O-Water exchange in photosystem II: Substrate binding and intermediates of the water splitting cycle. *Coord. Chem. Rev.* **252**, 306–317 (2008).
- Ugur, I., Rutherford, A. W. & Kaila, V. R. Redox-coupled substrate water reorganization in the active site of photosystem II-The role of calcium in substrate water delivery. *Biochim. Biophys. Acta* **1857**, 740–748 (2016).
- Suga, M. *et al.* Native structure of photosystem II at 1.95 Å resolution viewed by femtosecond X-ray pulses. *Nature* **517**, 99–103 (2015).
- Fraser, J. S. *et al.* Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl Acad. Sci. USA* **108**, 16247–16252 (2011).
- Tilton, R. F., Jr, Dewan, J. C. & Petsko, G. A. Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320 K. *Biochemistry* **31**, 2469–2481 (1992).
- Boussac, A., Rutherford, A. W. & Styring, S. Interaction of ammonia with the water splitting enzyme of photosystem II. *Biochemistry* **29**, 24–32 (1990).

11. Britt, R. D., Zimmermann, J. L., Sauer, K. & Klein, M. P. The state of manganese in the photosynthetic apparatus. 10. Ammonia binds to the catalytic Mn of the oxygen-evolving complex of photosystem-II - evidence by electron-spin echo envelope modulation spectroscopy. *J. Am. Chem. Soc.* **111**, 3522–3532 (1989).
12. Pérez Navarro, M. *et al.* Ammonia binding to the oxygen-evolving complex of photosystem II identifies the solvent-exchangeable oxygen bridge (μ -oxo) of the manganese tetramer. *Proc. Natl Acad. Sci. USA* **110**, 15561–15566 (2013).
13. Oyala, P. H., Stich, T. A., Debus, R. J. & Britt, R. D. Ammonia binds to the dangler manganese of the photosystem II oxygen-evolving complex. *J. Am. Chem. Soc.* **137**, 8829–8837 (2015).
14. Kupitz, C. *et al.* Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature* **513**, 261–265 (2014).
15. Sauter, N. K. *et al.* No observable conformational changes in PSII. *Nature* **533**, E1–E2 (2016).
16. Hellmich, J. *et al.* Native-like photosystem II superstructure at 2.44 Å resolution through detergent extraction from the protein crystal. *Structure* **22**, 1607–1615 (2014).
17. Moser, C. C., Keske, J. M., Warncke, K., Farid, R. S. & Dutton, P. L. Nature of biological electron transfer. *Nature* **355**, 796–802 (1992).
18. Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* **437**, 55–75 (1948).
19. Umena, Y., Kawakami, K., Shen, J.-R. & Kamiya, N. Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 Å. *Nature* **473**, 55–60 (2011).
20. Cox, N. *et al.* Electronic structure of the oxygen-evolving complex in photosystem II prior to O–O bond formation. *Science* **345**, 804–808 (2014).
21. Siegbahn, P. E. M. Structures and energetics for O₂ formation in photosystem II. *Acc. Chem. Res.* **42**, 1871–1880 (2009).
22. Glöckner, C. *et al.* Structural changes of the oxygen-evolving complex in photosystem II during the catalytic cycle. *J. Biol. Chem.* **288**, 22607–22620 (2013).
23. Li, X. & Siegbahn, P. E. M. Alternative mechanisms for O₂ release and O–O bond formation in the oxygen evolving complex of photosystem II. *Phys. Chem. Chem. Phys.* **17**, 12168–12174 (2015).
24. Kern, J. *et al.* Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. *Nat. Commun.* **5**, 4371 (2014).
25. Beck, W. F., Depaula, J. C. & Brudvig, G. W. Ammonia binds to the manganese site of the O₂-evolving complex of photosystem II in the S₂ state. *J. Am. Chem. Soc.* **108**, 4018–4022 (1986).
26. Hou, L. H., Wu, C. M., Huang, H. H. & Chu, H. A. Effects of ammonia on the structure of the oxygen-evolving complex in photosystem II as revealed by light-induced FTIR difference spectroscopy. *Biochemistry* **50**, 9248–9254 (2011).
27. Askerka, M., Vinyard, D. J., Brudvig, G. W. & Batista, V. S. NH₃ binding to the S₂ state of the O₂-evolving complex of photosystem II: Analogue to H₂O binding during the S₂ → S₃ transition. *Biochemistry* **54**, 5783–5786 (2015).
28. Retegan, M. *et al.* A five-coordinate Mn(IV) intermediate in biological water oxidation: spectroscopic signature and a pivot mechanism for water binding. *Chem. Sci. (Camb.)* **7**, 72–84 (2016).
29. Pecoraro, V. L., Baldwin, M. J., Caudle, M. T., Hsieh, W. Y. & Law, N. A. A proposal for water oxidation in photosystem II. *Pure Appl. Chem.* **70**, 925–929 (1998).
30. Vrettos, J. S., Limburg, J. & Brudvig, G. W. Mechanism of photosynthetic water oxidation: combining biophysical studies of photosystem II with inorganic model chemistry. *Biochim. Biophys. Acta* **1503**, 229–245 (2001).
31. Yamanaka, S. *et al.* Possible mechanisms for the O–O bond formation in oxygen evolution reaction at the CaMn₄O₅(H₂O)₄ cluster of PSII refined to 1.9 Å X-ray resolution. *Chem. Phys. Lett.* **511**, 138–145 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Director, Office of Science, Office of Basic Energy Sciences (OBES), Division of Chemical Sciences, Geosciences, and Biosciences (CSGB) of the Department of Energy (DOE) (J.Y., V.K.Y.) for X-ray methodology and instrumentation; National Institutes

of Health (NIH) grants GM055302 (V.K.Y.) for PS II biochemistry, structure and mechanism, GM110501 (J.Y.) for instrumentation development for XFEL experiments, GM102520 and GM117126 (N.K.S.) for development of computational protocols for XFEL data; the Ruth L. Kirschstein National Research Service Award (GM116423-02, F.D.F.); and the Human Frontiers Science Project Award No. RGP0063/2013 310 (J.Y., U.B., P.W., A.Z.). The DFG-Cluster of Excellence “UniCat” coordinated by T.U. Berlin and Sfb1078 (Humboldt Universität Berlin), TP A5 (A.Z., H.D.), the Solar Fuels Strong Research Environment (Umeå University), the Artificial Leaf Project (K&A Wallenberg Foundation 2011.0055) and Energimyndigheten (36648-1) (J.M.) are acknowledged for support. H.L. and C.A.S. acknowledge support from the US DOE, OBES, CSGB Division. W.I.W. and A.T.B. acknowledge support from an HHMI Collaborative Innovation Award. D.G.W. is funded by industrial income received by CCP4. This research used resources of NERSC, a User Facility supported by the Office of Science, DOE, under Contract No. DE-AC02-05CH11231. Portions of this work were supported by a BNL/US DOE, LDRD grant (11-008; A.M.O.); and NIH/NCRR grant 2-P41-RR012408, NIH/NIGMS grants 8P41GM103473-16 and P41GM111244 and the US DOE, OBER grant FWP BO-70 (A.M.O., B.A.). A.M.O. and P.T.D. were supported in part by the Diamond Light Source, and A.M.O. acknowledges support from a Strategic Award from the Wellcome Trust and the Biotechnology and Biological Sciences Research Council (grant 102593). P.B. was supported by a Wellcome Trust DPhil studentship. Testing of crystals and various parts of the setup were carried out at synchrotron facilities that were provided by the Advanced Light Source (ALS) in Berkeley and Stanford Synchrotron Radiation Lightsources (SSRL) in Stanford, funded by DOE OBES under contract DE-AC02-05CH11231 (ALS) and DE-AC02-76SF00515 (SSRL). The SSRL Structural Molecular Biology Program is supported by the DOE OBER and by the NIH (P41GM103393). Use of the LCLS and SSRL, SLAC National Accelerator Laboratory, is supported by the US DOE, Office of Science, OBES under Contract No. DE-AC02-76SF00515. We thank M. Bommer for discussions and help regarding structure refinement, crystallographic model building and validation, J. Hattné for his contributions to the development of XFEL diffraction data processing, A. Boussac for discussions on ammonia binding and his contributions to the substrate water exchange measurements of the S₃ state in the presence of ammonia, and the previous CXI beamline scientist, G. Williams, for his support during the initial stages of this project. We thank the support staff at LCLS/SLAC and at SSRL (BL 6-2, 7-3) and ALS (BL 5.01, 5.0.2, 8.2.1).

Author Contributions U.B., V.K.Y. and J.Y. conceived the experiment; R.A.-M., S.B., A.Z., J.M., U.B., N.K.S., J.K., V.K.Y. and J.Y. designed the experiment; I.D.Y., M.I., R.C., R.T., M.A.B., R.H., M.Z., L.D., I.S., A.Z. and J.K. prepared samples; M.S.H., A.A., J.E.K., J.R., M.L. and S.B. operated the CXI instrument; R.A.-M., T.J.L., J.E.K., J.R., M.L. and S.B. operated the MFX instrument; R.A.-M., J.M.G., S.N., M.S. and D.Z. operated the XPP instrument; S.G., S.K., F.D.F., H.L., E.P., B.A., A.M.O., R.G.S., C.A.S., C.S., J.M. and J.K. developed, tested and ran the sample delivery system; R.C., S.K., C.d.L., L.V.P., H.N., M.H.C., D.Sh., J.M. and J.Y. performed and analysed O₂ evolution and EPR measurements; I.D.Y., M.I., R.C., S.G., S.K., A.S.B., R.A.-M., F.D.F., T.K., T.M.-C., H.L., R.G.S., C.A.S., R.H., M.Z., L.D., M.K., C.d.L., C.S., D.So., T.-C.W., E.P., C.W., T.F., P.A., P.B., B.A., P.T.D., A.M.O., J.M.G., S.N., M.S., D.Z., M.S.H., T.J.L., A.A., J.E.K., J.R., M.L., S.B., P.W., A.Z., J.M., U.B., N.K.S., J.K., V.K.Y. and J.Y. performed the LCLS experiment; I.D.Y., A.S.B., T.M.-C., A.Y.L., M.U., N.W.M., D.L., P.V.A., D.G.W., G.E., W.I.W., A.T.B., P.H.Z., P.D.A. and N.K.S. developed new software for data processing; I.D.Y., A.S.B., F.D.F., C.W., T.F., L.L., P.A., P.B., T.K., T.M.-C., H.D., N.K.S. and J.K. processed and analysed XFEL data; I.D.Y., R.C., J.M., J.K., J.Y. and V.K.Y. wrote the manuscript with input from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.Z. (athina.zouni@hu-berlin.de), J.M. (johannes.messinger@umu.se), V.K.Y. (vkyachandra@lbl.gov) or J.Y. (jjano@lbl.gov).

Reviewer Information *Nature* thanks J. Murray, C. Yocum and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Sample preparation. PS II dimers from *Thermosynechococcus* (*T.*) *elongatus* were prepared using a modification of the protocol described in ref. 32, substituting the C₁₂E₈ detergent for β -DM and using betaine as a cryoprotectant instead of glycerol¹⁶. Crystallization was achieved by addition of PEG 5000 and seeding methods³³ to obtain a uniform size distribution of 5–15 μ m for the MESH (microfluidic electrokinetic sample holder) injector (CXI instrument at LCLS) and 20–50 μ m for the ADE (acoustic droplet ejection)-DOT (droplet on tape) injector (XPP and MFX instruments at LCLS) (see Sample Injection and Illumination). Crystals were dehydrated by treatment with high concentrations of PEG 5000 before measurement. The final buffer used for the XRD measurements for the CXI instrument was 100 mM TRIS-Cl, pH 7.5, 100 mM (NH₄)₂SO₄, 31.5% (w/v) PEG 5000, 10% ethylene glycol. This procedure resulted in a change in unit cell parameters of the crystals compared to earlier crystallization protocols due to differences in crystal packing. The packing obtained from this procedure¹⁶ is very similar to the arrangement of PS II dimers in the native thylakoid membrane of *T. elongatus*. The final buffer used for the XRD measurements for the XPP and MFX instruments was 100 mM MES pH 6.5, 100 mM ammonium chloride, 10% ethylene glycol and 35% (w/v) PEG 5000. Crystalline unit cells for structures determined at the CXI instrument (S₁ and 2F-NH₃) differed from the one determined at the XPP and MFX instruments (2F), particularly in the *c*-axis dimension (Extended Data Table 1). The differences in the unit cell parameters are due to the different buffer conditions and dehydration procedures. Nevertheless, a similar dimer–dimer interaction in the direction of the membrane plane was observed in both crystal forms.

Characterization of ammonia-treated PS II and establishing the preparation protocol. Prolonged incubation of PS II samples at higher pH, and with ammonia and TRIS, are known to reduce the PS II oxygen evolution activity. For this reason, thorough control experiments were undertaken to characterize the light-induced turnover of purified PS II dimers under crystallization conditions in comparison to conditions at lower pH and without ammonia or TRIS addition. The buffers used were A (100 mM MES, pH 6.0, 10 mM CaCl₂, 30% glycerol) for low pH conditions, B (100 mM TRIS-Cl, pH 7.5, 100 mM (NH₄)₂SO₄, 30% glycerol) to mimic crystallization conditions in solution samples, C (100 mM TRIS-Cl, pH 7.5, 100 mM (NH₄)₂SO₄, 31.5% PEG 5000, 10% ethylene glycol) for measurements on crystals and D (100 mM TRIS-Cl, pH 7.5, 10 mM CaCl₂, 30% glycerol) for measurements at high pH without ammonia. EPR measurement of the content of Mn²⁺ released from crystals after 12 h incubation in buffer C yielded an upper bound of 5%.

O₂ activity. Measurement of the O₂ yield by means of a Clark-type electrode from ammonia-treated crystals under continuous illumination showed 70–75% of the activity observed under low pH conditions (pH 6.0). To see whether this reduction in activity is due to the pH effect or to some fraction of Cl[−] binding sites being substituted with ammonia, we compared the O₂ activity of PS II in buffer A (pH 6.0, no ammonia), in buffer B (pH 7.5 with ammonia), and in buffer D (pH 7.5, CaCl₂, no ammonia). The O₂ evolution rates of PS II in buffer B and in buffer D were similar and were 70% of the activity in buffer A ($\sim 3,000 \mu\text{mol O}_2/(\text{mg}(\text{Chl}) \times \text{h})$) with PPBQ and $3,600 \mu\text{mol O}_2/(\text{mg}(\text{Chl}) \times \text{h})$ with DCBQ. Therefore, we conclude that the reduction in the O₂ activity in our ammonia-treated PS II is an effect of the higher pH, not of the displacement of chloride. Similar results were obtained by Joliot-type O₂ evolution measurements (no electron acceptors added, 2 Hz flash frequency, Xe-flash lamp) on the same type of PS II core preparations (Extended Data Fig. 7f–i). These experiments revealed a similar O₂ oscillation pattern and comparable total O₂ yields with and without ammonia or TRIS or MOPS buffer at pH 7.5, even if incubated for several hours. One exception was that the total O₂ yield of the sample in buffer containing 100 mM (NH₄)₂SO₄ was 30% smaller than when suspended in buffer containing 100 mM Na₂SO₄. The data also show that our core preparations have a large enough plastoquinone pool to allow a full cycle of O₂ production in the absence of added electron acceptors (Extended Data Fig. 7a). This was further confirmed by membrane-inlet mass spectrometry experiments in which at very low frequency (12 s between flashes) an increase in O₂ yield of only 15% was observed if PPBQ was added (Extended Data Fig. 7b). The experiments were performed and analysed as described earlier³⁴.

Importantly, nearly full activity (O₂ rates) can be restored in solution samples by exchanging the sample back into pH 6.5 buffer, but keeping the (NH₄)₂SO₄ concentration constant. This implies that the OEC remains intact and that the lower overall activity at the higher pH may be associated with changes in the protonation state of residues involved in proton transfer networks around the OEC.

Electron density of the Cl[−] binding site. As described above, ammonia does not bind to the Cl[−] binding site(s) in the presence of excess Cl[−] (100 mM) in PS II, and the O₂ activity results support this conclusion. To further confirm this, we checked the occupation of the two Cl[−] binding sites in the electron density map of the 2F data. As shown in Extended Data Fig. 9a, b, both sites are occupied by chloride. We have also checked this by substituting chloride with ammonia in the

structural model and calculating electron density difference maps using this substituted model after three cycles of refinement in *phenix.refine*. The appearance of positive difference density upon substitution confirmed that the density observed at both chloride-binding sites in the electron density maps from our dark and 2F data cannot be explained by ammonia (Extended Data Fig. 9a, b). In summary, we can conclude that ammonia does not bind to the chloride-binding sites under the current experimental conditions.

EPR spectra of ammonia-treated PS II. Previous studies have suggested that ammonia binds upon formation of the S₂-state and stays bound in the S₂–S₃ transition¹⁰. Therefore, we used the well-characterized S₂ EPR multiline signal (MLS) to infer the extent of ammonia binding in our 2F samples. The binding of ammonia to Mn in the S₂-state was confirmed by the altered EPR multiline spectrum. We measured EPR from both single flash (data not shown) and continuous illumination conditions with and without annealing to populate the S₂-state. These different procedures yielded similar results and for comparison with published data we show here the spectra obtained using continuous illumination and annealing (Extended Data Fig. 7d, e). We compared the native PS II EPR multiline spectrum of the S₂-state (pH 6.0, no ammonia), the native PS II EPR multiline spectrum of the S₂-state (pH 7.5, no ammonia), and the ammonia-treated S₂-state spectrum (pH 7.5 in the presence of 100 mM ammonia). The spectral changes observed in the ammonia-treated S₂-state sample were similar to results reported previously^{13,35} and the altered MLS was observed. Comparing spectral heights between the altered and native MLS is challenging as the peaks do not clearly correspond to each other. Nevertheless, when we assume that the MLS intensity of normal PS II (pH 7.5 without ammonia) and the altered MLS of ammonia-bound PS II (pH 7.5 with ammonia) both similarly reflect the number of unpaired spins, we estimate the altered (ammonia-treated) MLS intensity to be $\sim 70\%$ of that of the non-ammonia-treated PS II (calculated by the averaged peak heights of multiple EPR peaks at the high field side of the Y_D signal). The reduction in the MLS gives a lower limit of 70% for centres in the S₂-state as it could also be that, under elevated pH conditions, a fraction of the centres in the S₂ state do not give an MLS³⁶.

Determination of the S₃ state population in the native and ammonia-treated PS II 2F states. Flash-induced oxygen measurements in a replica of the capillary setup used in the XRD experiment were performed using membrane inlet mass spectrometry (MIMS) for O₂ detection^{24,37,38}. Analysis of the flash data obtained for PS II crystals incubated for 12 h in buffer C (Extended Data Fig. 7f) showed that 50–60% of the centres that are active in oxygen evolution are in the S₃-state after two light flashes in the ammonia-treated PS II. We therefore conclude that in the 2F ammonia-treated PS II samples, S₃ is formed in 50–60% of the centres, with most of the remaining centres ($\sim 25\%$) being in the S₂-state. Therefore, 75–85% of the centres are expected to bind ammonia in the 2F samples. This estimate was confirmed with the independent Joliot-type flash-induced oxygen evolution measurements on solution samples of our PS II core complexes. The analysis of these patterns (Extended Data Fig. 7g–i) gives a S₃ state population of about 53% in the 2F samples at pH 7.5 (TRIS, (NH₄)₂SO₄), and of 60% at pH 6.5 (MES, (NH₄)₂SO₄) (Extended Data Fig. 7a, g–i).

At pH 7.5, the O₂ evolution rate is reversibly reduced to 70–75% of its maximum value at pH 6.0, and the MLS amplitude is reduced to a similar extent (Extended Data Fig. 7d). Similarly, the total O₂ yield obtained in the flash-induced oxygen evolution pattern was about 50% at pH 7.6 compared to pH 6.5. As we do not know with certainty whether these reductions are due to reversible blockage of centres in a particular S-state or due to kinetic limitations (O₂) or changes in the hyperfine couplings (EPR)³⁶, some uncertainty arises as to the S-state population of the 2F-NH₃ samples when normalized to all centres. Therefore, the most conservative estimate for the S-state distribution in the current 2F-NH₃ crystals is 25% S₃-state. In contrast, the highest estimate of the S₃-state occupancy in the 2F-NH₃ sample becomes 50%, with 25% in the S₂-state.

Sample injection and illumination. Crystals in high PEG 5000 buffer (31.5% w/v) were injected using a modified version of the electrospinning injector (MESH) from ref. 39. In the modified version, a double capillary setup was used, allowing the protection of the crystals in mother liquor with a shield flow of 50% ethylene glycol. The setup is discussed in detail in ref. 40. Illumination of samples was performed as described previously²⁴ and optimal illumination parameters were established by parallel oxygen yield measurements using MIMS (Extended Data Fig. 7b, c, f). The experimental setup at the CXI instrument^{41,42} of LCLS was similar to the one used in our previous work^{24,43}. For the ammonia-treated doubly illuminated (2F-NH₃) data, each volume segment of the crystal suspension was illuminated by 120-ns laser pulses ($20 \pm 2 \text{ mJ/cm}^2$) at 527 nm from lasers 2 and 3 along the sample delivery capillary, resulting in a delay time of 0.5 s between the first and second illuminations and of 0.5 s between the second illumination and the X-ray probe.

For the high-resolution 2F structure of the native PS II, the data were collected at the XPP⁴⁴ and the MFX⁴⁵ instruments of LCLS. The newly developed DOT sample

delivery method was used in combination with an ADE method (Fuller, F. D. *et al.*, submitted). The laser illumination timing remains the same as that used for MESH with the difference of 1 s spacing between first and second fibre excitations and between the last fibre excitation and the X-ray probe, using 100-ns laser pulses at 527 nm from a Nd:YLF laser (Evolution, Coherent). To test light saturation for the DOT system, a 100–150- μm film was established with the help of a washer between the silicon membrane of the mass spectrometer inlet and a thin microscope glass plate (thin layer MIMS setup). In this test setup the samples were illuminated with a pulsed Nd:YAG laser (532 nm, Continuum). The laser energy was measured at the sample position. To allow resolution of the individual O_2 yields (Y_n ; n = flash number), the flashes had to be spaced by 12 s, which leads to a significantly lower Y3/Y4 ratio (Extended Data Fig. 7b) than at 1 or 2 Hz as used in the Joliot-type experiments (Extended data Fig. 7a, g–i) and during the XFEL measurements. The data in Extended Data Fig. 7c show that the samples are saturated at 70 mJ/cm². At the XFEL a light intensity of 120 ± 10 mJ/cm² was applied.

X-ray diffraction setup and data processing. PS II XRD data collected at the CXI instrument^{41,42} of LCLS were recorded on a CSPAD detector⁴⁶ operating at a frame rate of 120 Hz over an aggregate total time period of 641 min and processed using *cctbx.xfel*^{47,48}. To avoid saturating pixels on the CSPAD, which has a limited dynamic range of ~ 350 photons per pixel at 8 keV in its high gain mode⁴⁶, we used the CSPAD detector in a mixed gain mode, putting the low resolution pixels in a low-gain setting while the high resolution pixels were set to high gain. The low-gain mode is less sensitive to low signal but harder to saturate, thus preserving bright, low resolution reflections. Conversely, the high-gain setting is easier to saturate, but is more sensitive to low signal, as is typical for high resolution reflections. After subtracting a pedestal estimate derived from an uncorrected, dark average, we applied a gain multiplier of 6.88 to the pixels in the low-gain setting, thereby putting low- and high-gain pixels on a similar scale. This number is merely an estimate based on matching the background levels at the low–high gain boundary. A more thorough exploration of gain for protein diffraction data on the CSPAD detector is planned.

Thermolysin pseudo-powder patterns were generated by taking the maximum value of each pixel from the ensemble of diffraction patterns in a reference thermolysin dataset⁴⁰ collected at a known detector distance. A precise sample-to-detector distance and the locations and orientations of the 64 sensors on the CSPAD X-ray diffraction detector were obtained by refining the geometries of the 64 sensors against all the single crystal models from the dataset⁴⁸.

PS II XRD data collected at the XPP⁴⁴ and the MFX⁴⁵ instruments were recorded on a Rayonix MX170-HS detector operating at its maximum frame rate of 10 Hz in the 2-by-2 binning mode, in which square bins of four pixels are configured to share underlying electronics and act as a single pixel. This mode was selected as a compromise between high data acquisition rates at larger effective pixel sizes and improved ability to resolve individual Bragg spots, especially for large unit cells, at smaller effective pixel sizes. In all binning modes, the dynamic range of the Rayonix detector is also larger than that of the CSPAD. Using this detector and the ADE–DOT sample delivery system, XRD data could be acquired from larger crystals (20–50 μm) than were used at the CXI instrument using the CSPAD and the modified MESH sample delivery system (5–15 μm).

Images were indexed using the Rossmann algorithm^{49,50} as implemented in *LABELIT*⁵¹, with the choice of lattice basis guided by a target unit cell of $a = 118.2$ Å, $b = 224.6$ Å, $c = 331.9$ Å, $\alpha = \beta = \gamma = 90^\circ$ (later refined for data merging, Extended Data Table 1) for the dark and 2F-NH₃ PS II datasets. This cell was determined by examining the distribution of unit cell dimensions found when indexing with no target unit cell. Indexing was attempted on all frames, without a pre-filtering step, so as to obtain the maximal number of indexed images. As previously described in detail⁵² the Rossmann algorithm produces three basis vectors that generate a primitive triclinic lattice. Miller indices are then deduced for the strong spots, and the lattice parameters (unit cell and crystal orientation) are refined against the observed spot positions. Next, the lattice parameters are constrained to the known orthorhombic symmetry of the space group (in this case, $P2_12_12_1$), meaning the cell angles are all set to 90° , and the remaining free lattice parameters are re-refined. However, we noticed a potential problem with applying the orthorhombic constraints. In the earliest trials, we redetermined the Miller indices of the strong spots based on their proximity to lattice nodes, after applying the 90° constraints. We found that this can sometimes assign Miller indices that are misindexed by one unit along the c -axis, since the large ~ 332 Å cell length gives lattice nodes that are very close together. We therefore incorporated a new option within *cctbx.xfel* to apply high-symmetry constraints, but skip the step of redetermining the Miller indices; instead, the initially determined Miller indices were converted from the triclinic to the high-symmetry setting using a change-of-basis operator⁵³ before performing the final round of parameter refinement. Refinement was implemented using the newly introduced *DIALS* toolkit⁵⁴, which implements

a target function based on both the observed spot positions and the reciprocal lattice points' angular proximity to the Ewald sphere⁵², while also permitting the refinement of additional parameters such as the detector tilt. As in previous work⁴⁸, strong spots not covered by the modelled lattice were considered separately in an attempt to index a second lattice on each image for the dark and 2F-NH₃ datasets. We note that different unit cell parameters were obtained for the higher resolution 2F data of native PS II collected at the XPP and the MFX instruments. Using the ADE–DOT method (Fuller, F. D. *et al.*, submitted) and slightly different dehydration conditions, an average unit cell of $a = 117.9$ Å, $b = 223.1$ Å, $c = 310.7$ Å, $\alpha = \beta = \gamma = 90^\circ$ was obtained, and a subsequent indexing step using these parameters as constraints was used to generate the final native PS II 2F dataset.

For the integration and merging of high-resolution reflections, we took into account various community concerns that weak reflections ought to be included since they contain measurable information⁵⁵, while at the same time realizing that shot-to-shot and crystal-to-crystal variation requires differing resolution cutoffs for each integrated image. For the work described here, the following solution was adopted: for each image, a first round of spot prediction and integration was performed, with a high-resolution cutoff based on the apparent limit of bright spots from the spotfinding program⁵⁶. The intensity/standard deviation ratio ($I/\sigma(I)$) was examined as a function of diffraction angle to determine the resolution bin where $\langle I/\sigma(I) \rangle = 0.5$, essentially identifying a zone within which the signal-to-noise falls to a low, but non-zero value. Then, for a second round of spot prediction and integration, the limiting resolution was set to a value far beyond the $\langle I/\sigma(I) \rangle = 0.5$ limit, in order to ensure that any reasonable positive signal was integrated. This 'greedy' integration limit was set to a reciprocal-spacing value that encloses 1.5 times as much reciprocal-space volume as the limit determined by $\langle I/\sigma(I) \rangle = 0.5$.

Based on the integrated intensity measurements from this second round of integration, individual resolution limits were then determined for each image as follows: $\langle I/\sigma(I) \rangle$ values were computed in resolution bins as a function of the diffraction angle, and the bins were excluded from further data merging beginning in the bin where $\langle I/\sigma(I) \rangle$ first falls below 0.1. This threshold, which is essentially zero, implies that measurements beyond the limiting bin are half positive and half negative on average, suggesting that no actual Bragg signal is present.

Before scaling and merging the dark and 2F-NH₃ datasets, images with a refined c axis that was not between 325 and 340 Å were discarded. In the case of the native 2F dataset, the unit cell constraint during indexing ensured that no images with aberrant unit cells were integrated. Additionally, we examined the distribution of refined beam centres for all images. The beam centre (indicating the relation between the detector and the incident X-rays) is freely refined during parameter refinement, even though the true relative positions of the beam and detector are fixed to within a micrometre. The small fraction of images whose beam centre differed by more than one pixel from the mean were discarded, as these lattices were mis-indexed by one Miller index unit along a crystallographic axis, usually the long c axis. As the unit cell lengths of this crystal form differ from those of other PS II crystal forms reported elsewhere¹⁶, reference intensities for scaling were not available. We therefore merged and scaled the data initially without an external reference, using our postrefinement program *PRIME* recently described⁵⁷. A resolution cutoff of 2.5 Å was chosen in this step to match the apparent resolution limit of the raw diffraction patterns. A total of 1,264, 22,311 and 2,294 diffraction images were included in the S₁-state, native 2F and the 2F-NH₃ datasets, respectively (Supplementary Table 1). Following merging and post-refinement, complete datasets for the dark, native 2F and 2F-NH₃ states were obtained at 3.0, 2.25 and 2.8 Å resolution, respectively, and structural models were refined (see below for detailed description) to $R_{\text{work}}/R_{\text{free}}$ of 0.2637/0.3030 for the dark, 0.1949/0.2308 for the 2F and 0.2497/0.2997 for the 2F-NH₃ state (Extended Data Table 1, Supplementary Tables 2–4).

For the dark and 2F-NH₃ datasets an initial model was determined with *Phaser* molecular replacement⁵⁸ using two copies of a hybrid model of the PS II monomer composed of the cryogenic S₁-state PS II structure from *T. vulcanus* (PDB ID 4UB6⁷) and the additional chain present in native PS II from *T. elongatus* (PDB ID 4PJ0¹⁶). The initial model was refined using *phenix.refine*⁵⁹ and used to calculate model structure factors, which were then used as a scaling reference to re-scale the original data using the program *cximerge*. For the native 2F dataset the 2F-NH₃ model was used for *Phaser* molecular replacement. Examination of the $I/\sigma(I)$ and completeness measures as a function of resolution allowed us to select new resolution cutoffs at this stage and to re-merge the data with *cximerge* to these cutoffs. Negative measurements were included during merging instead of being discarded as had been done previously⁴⁸. Generally, including negative measurements moves the distribution of merged intensities closer to that expected for crystallographic data, as measured using the L and Z tests⁶⁰ (results not shown). Further investigation of this effect is ongoing and will be the subject of a future work.

Custom restraints for the Mn_4CaO_5 complex geometry for the dark state complex, derived from ref. 7, were generated by averaging bond lengths and angles over the four crystallographically independent monomers in two structures (PDB IDs 4UB6 and 4UB8) and imposing custom bond length and angle tolerances (see Supplementary Table 7 for restraints used). Average nonbonding distances between the cluster and surrounding residues were also calculated across the four monomers and adapted as restraints for the room temperature structures. Water molecules directly coordinating cluster metal atoms were restrained to match those in ref. 7 in early cycles of refinement, and these restraints were subsequently modified to minimize difference density, resulting in separate sets of metal–water restraints for the two crystallographically nonequivalent PS II monomers. Custom restraints were also generated for the α - and β -pucker chlorophyll-*a* ligands in order to effectively restrain the planarity of the porphyrin ring, the magnitude of displacement of the Mg centre from the plane of the ring, the direction of this displacement relative to the phytol tail (which differs between α - and β -pucker stereoisomers), and the Mg–His or Mg–water coordination distances at this medium resolution. The chlorophyll restraints were based on the default chlorophyll-*a* (CLA) ligand CIF file distributed with *Phenix*⁶¹. In the higher-resolution native 2F dataset, the data quality was sufficient to generalize to a single chlorophyll-*a* restraints file for both α - and β -puckers, but restraints maintaining coordination of chlorophyll magnesium atoms with nearby histidine sidechains or water molecules were still necessary to override the automatic repulsion of atoms within van der Waals distance of each other.

Model building and map calculation. Model building was performed in *Coot*⁶² and figures were generated using *PyMol* (Schrödinger, LLC). Model building was aided by recently-developed tools: feature-enhanced maps⁶³, designed to scale all non-solvent density to a uniform level, were used to identify highly flexible portions of ligands and detergent molecules once refinement had approached convergence⁵⁹. Polder omit maps (Liebschner *et al.*, submitted), a form of omit map newly available with the *Phenix* package, were used to test the contribution of model bias to the observed electron density at selected ligands and TMHs and to identify unmodelled water molecules in the vicinity of the Mn_4CaO_5 cluster. Polder omit maps are able to reveal weaker features than traditional omit maps by uniformly omitting bulk solvent from the omitted region and its surroundings, a key advantage when locating ordered solvent. To test model bias, polder omit maps were calculated after perturbing the model by omitting an area of interest: the selected ligand or residues were omitted, the resulting model was subjected to three cycles of coordinate and real space refinement in *phenix.refine*, the omitted ligand or residues were re-inserted (to allow identification of the region from which to omit bulk solvent), and the polder omit map was calculated (once again omitting the selected atoms in addition to the surrounding bulk solvent). Simulated annealing polder omit maps were calculated identically but with simulated annealing enabled during coordinate refinement. A comparison of normal $mF_o - DF_c$ and polder omit maps with real space or simulated annealing is shown in Extended Data Fig. 2e–h. The advantage of the polder compared to the $mF_o - DF_c$ map is manifested in lower noise and higher levels of detail of the electron density maps (compare Extended Data Fig. 2g, h). Comparing simulated annealing and real space polder omit maps for several different regions in the PS II complex (for example, Extended Data Fig. 2f, h) shows that both are very similar, with the simulated annealing being slightly more disruptive to the structure. As no additional benefit of the simulated annealing protocol was found, all other polder omit maps shown in this work were generated after real space refinement of the omit model.

Modelling of waters. Waters were incorporated into the model in two different steps. After initial refinement of the model, waters were placed into the $2mF_o - DF_c$ map using the *Phenix* auto water placement option during several subsequent coordinate refinement cycles. These positions were manually checked in *Coot*, and waters with strong enough electron density and good hydrogen bonding environments were then included in subsequent runs of refinement of the model with auto water placement disabled and coordinate and real space refinement enabled, resulting in 124, 1,179 and 107 ordered waters in the dark, 2F and 2F-NH₃ datasets, respectively (Extended Data Fig. 5). Upon convergence, polder omit maps excluding all waters were generated, and these were inspected for the placement of possible additional waters in the region around the OEC and within hydrogen bonding distance from hydrogen bonding-capable residues in the final model. For the two lower resolution datasets, waters were placed manually into polder maps in *Coot* and their positions were fit to the density using the *Coot* rigid body fit tool. The resulting model was subsequently refined for three cycles in *phenix.refine* with coordinate and real space refinement enabled. This resulted in an additional 33 and 55 water positions for the dark and 2F-NH₃ datasets, respectively (see for example, Extended Data Fig. 5d). The coordinates

for these additional waters are not included in the deposited PDB files but are given in Supplementary Table 6.

Estimated positional precision. To estimate the error when calculating differences in distances between cofactors, we sought to identify a method for determining the coordinate error of a molecule or residue sequence as a unit, for which the maximum likelihood estimate for atomic coordinate error (0.5 Å for our dark and 2F-NH₃ models, 0.34 Å for our native 2F model, 0.34 Å for 4PJ0 and 0.27 Å for 4UB6) is not a good guide. Instead we estimated the error in the positions of larger segments or molecules by generating simulated annealing omit maps of individual chlorophylls and TMHs. Treating the omitted unit as a rigid body, we obtained the best fit of the unit to polder difference density and calculated the magnitude of the shift as the distance between the centre of the unit in the refined model and the centre of the unit placed in the difference density. In the case of chlorophylls, the centre was defined as the average of the positions of the four porphyrin nitrogens, and in the case of the TMHs, the average of the positions of all α -carbons was used. On the basis of these results, we estimate that we can position an entire Chl molecule with better than 0.13 Å precision in the dark model and 0.10 Å in the 2F and 2F-NH₃ models. Similarly, for a TMH, we arrive at an upper bound of 0.08 Å precision for all three models. It is expected that the positional error should approach the maximum likelihood estimate for atomic coordinate error as the size of the unit decreases. Using the same procedure, the non-haem Fe^{II} shifted by 0.5–0.8 Å in the dark model, 0.0 Å in the 2F model and 0.3–0.5 Å in the 2F-NH₃ model.

To estimate the precision of deriving the metal ion positions in the OEC, we compared a modification of the above technique and a difference-density generating technique. For the former, individual metal atoms were omitted from the cluster, but since polder maps are not applicable when omitting single atoms from a larger molecule, standard $mF_o - DF_c$ maps were used. Also, to prevent the collapse of coordinating waters into the cluster, waters coordinating Ca1 or Mn4 were omitted along with these atoms and restored to the annealed model before calculating $mF_o - DF_c$ maps. The mobile W4, which was placed after refinement and is not included in the deposited model for the dark and 2F-NH₃ datasets, was also added to the annealed model before map calculation for Ca1 only, since the difference density at the Ca1 centre was skewed in its absence. Observed shifts for each metal, averaged across the two monomers, ranged from 0.2 to 0.6 Å in the dark model, 0.1 to 0.6 Å in the 2F model and 0.2 to 0.4 Å in the 2F-NH₃ model. These values are likely to be overestimates owing to the propensity for the whole cluster to shift into the open density during refinement and for the surrounding, coordinated waters to affect the difference density.

For the difference-generating technique, we shifted individual metals in increments of 0.1 Å and calculated the resulting difference density ($mF_o - DF_c$). We propose that the magnitude of shift necessary to generate paired positive and negative difference density on either side of the atom can serve as an estimate of the positional error in OEC metal positions. A shift of Mn4 of 0.1 Å along the Mn4–Mn3 direction away from Mn3 led to paired difference density at the 2σ contour level and a shift of 0.2 Å led to strong paired difference density at the 2.5σ contour level in the lowest resolution dataset, clearly above the noise in the surrounding in the difference map. Combining the results from both approaches, we estimate the precision of our metal positions to be in the range of 0.2–0.4 Å for the dark model, 0.1–0.3 Å for the 2F model and 0.2–0.3 Å for the 2F-NH₃ model.

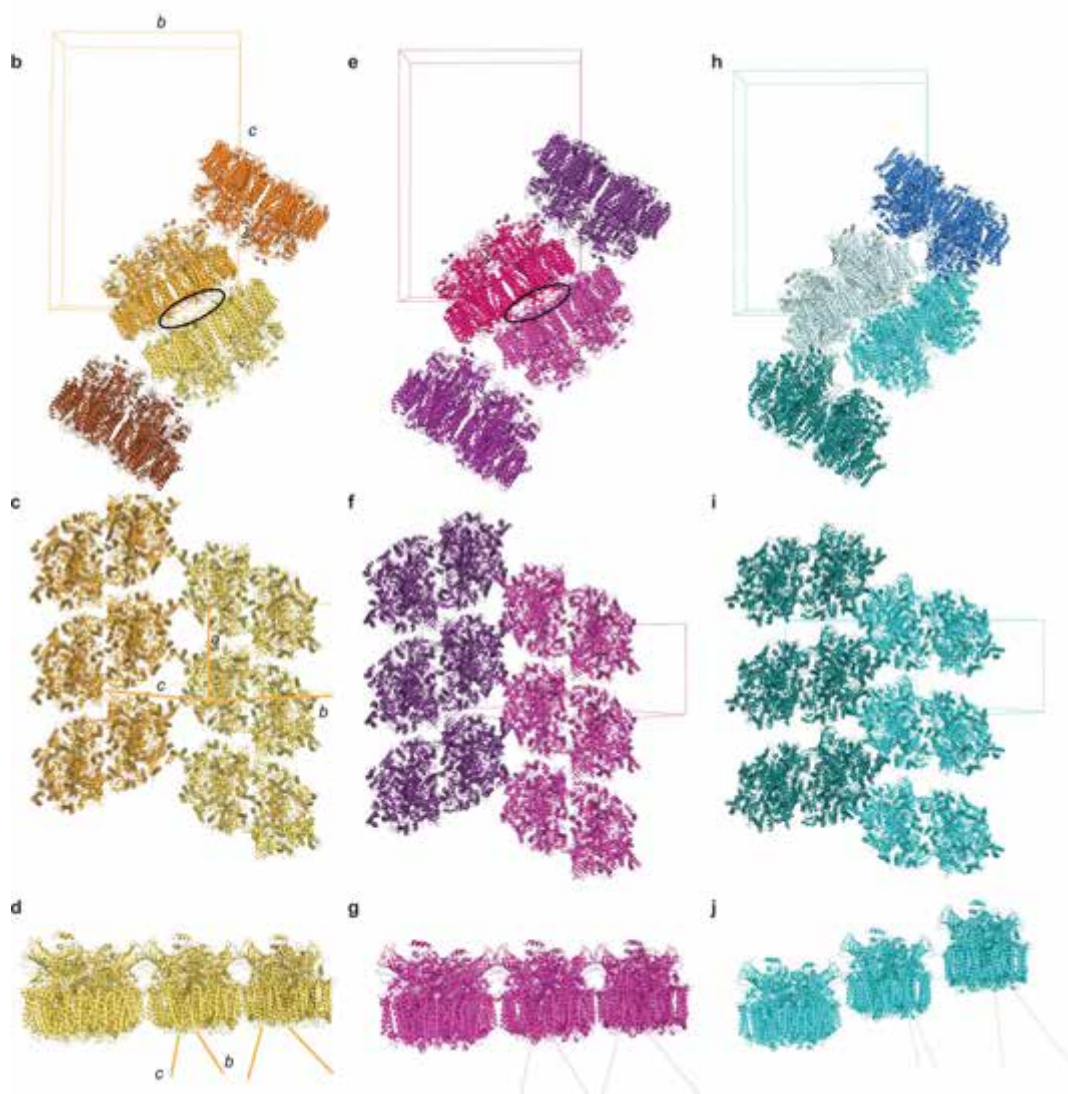
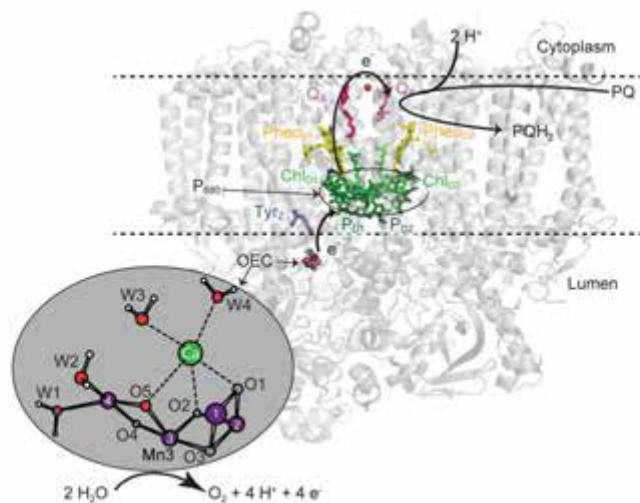
Comparison between the different PS II structures. Monomers or dimers of PS II from the 4UB6 and 4PJ0 models were superimposed onto our room temperature dark and 2F structural models using *Coot*. Cofactor distances were calculated as centre–centre distances. In the case of Chl and Pheo, the centre was defined as the average of the positions of the four porphyrin nitrogens. To compare distances of cofactors from the pseudo C2-axis, the axis was defined as the membrane normal passing through the middle between the Pheo_{D1} and Pheo_{D2} molecules and distances of each pigment centre from this axis were computed. For comparison of individual residue positions, the models were aligned for short windows of 5–20 residues. For each of the two datasets, the quality of match between the $2mF_o - DF_c$ maps (together with the refined models) and the 4PJ0 or the 4UB6 models was inspected visually and scored. For several residues clear difference density was visible in the initial rounds of model building when the starting model (based on the cryogenic structures) was used for generation of the electron density maps. Additional, polder $mF_o - DF_c$ maps were used to inspect different rotamer conformations.

Code availability. Links to the software described here (*phenix*, *DIALS*, *cctbx.xfel*) and to specific instructions for processing XFEL data are given at <http://ccilbl.gov>.

Data availability. The atomic coordinates and structure factors have been deposited in the Protein Data Bank under accession numbers 5KAF (dark), 5TIS (2F) and 5KAI (2F-NH₃).

32. Kern, J. *et al.* Purification, characterisation and crystallisation of photosystem II from *Thermosynechococcus elongatus* cultivated in a new type of photobioreactor. *Biochim. Biophys. Acta* **1706**, 147–157 (2005).
33. Ibrahim, M. *et al.* Improvements in serial femtosecond crystallography of photosystem II by optimizing crystal uniformity using microseeding procedures. *Struct. Dyn.* **2**, 041705 (2015).
34. Pham, L. V. & Messinger, J. Electrochemically produced hydrogen peroxide affects Joliot-type oxygen-evolution measurements of photosystem II. *Biochim. Biophys. Acta* **1837**, 1411–1416 (2014).
35. Beck, W. F. & Brudvig, G. W. Binding of amines to the O₂-evolving center of photosystem II. *Biochemistry* **25**, 6479–6486 (1986).
36. Geijer, P., Deák, Z. & Styring, S. Proton equilibria in the manganese cluster of photosystem II control the intensities of the S(0) and S(2) state $g \approx 2$ electron paramagnetic resonance signals. *Biochemistry* **39**, 6763–6772 (2000).
37. Yano, J. *et al.* in *Sustaining Life on Planet Earth: Metalloenzymes Mastering Dioxygen and Other Chewy Gases* Vol. 15 (eds Sosa Torres, M.E. & Kroneck, P.M.H.) 13–43 (Springer, 2015).
38. Beckmann, K., Messinger, J., Badger, M. R., Wydrzynski, T. & Hillier, W. On-line mass spectrometry: membrane inlet sampling. *Photosynth. Res.* **102**, 511–522 (2009).
39. Sierra, R. G. *et al.* Nanoflow electrospinning serial femtosecond crystallography. *Acta Crystallogr. D* **68**, 1584–1587 (2012).
40. Sierra, R. G. *et al.* Concentric-flow electrokinetic injector enables serial crystallography of ribosome and photosystem II. *Nat. Methods* **13**, 59–62 (2016).
41. Boutet, S. & Williams, G. J. The Coherent X-ray Imaging (CXI) instrument at the Linac Coherent Light Source (LCLS). *New J. Phys.* **12**, 035024 (2010).
42. Liang, M. *et al.* The Coherent X-ray Imaging instrument at the Linac Coherent Light Source. *J. Synchrotron Radiat.* **22**, 514–519 (2015).
43. Kern, J. *et al.* Simultaneous femtosecond X-ray spectroscopy and diffraction of photosystem II at room temperature. *Science* **340**, 491–495 (2013).
44. Chollet, M. *et al.* The X-ray Pump-Probe instrument at the Linac Coherent Light Source. *J. Synchrotron Radiat.* **22**, 503–507 (2015).
45. Boutet, S., Cohen, A. E. & Wakatsuki, S. The new macromolecular femtosecond crystallography (MX) instrument at LCLS. *Synchrotron Radiat. News* **29**, 23–28 (2016).
46. Herrmann, S. *et al.* CSPAD upgrades and CSPAD V1.5 at LCLS. *J. Phys.* **493**, 012013 (2014).
47. Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W. & Echols, N. New Python-based methods for data processing. *Acta Crystallogr. D* **69**, 1274–1282 (2013).
48. Hattne, J. *et al.* Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. *Nat. Methods* **11**, 545–548 (2014).
49. Steller, I., Bolotovskiy, R. & Rossmann, M. G. An algorithm for automatic indexing of oscillation images using Fourier analysis. *J. Appl. Crystallogr.* **30**, 1036–1040 (1997).
50. Rossmann, M. G. & van Beek, C. G. Data processing. *Acta Crystallogr. D* **55**, 1631–1640 (1999).
51. Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. Robust indexing for automatic data collection. *J. Appl. Crystallogr.* **37**, 399–409 (2004).
52. Sauter, N. K. *et al.* Improved crystal orientation and physical properties from single-shot XFEL stills. *Acta Crystallogr. D* **70**, 3299–3309 (2014).
53. Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. Improved statistics for determining the Patterson symmetry from unmerged diffraction intensities. *J. Appl. Crystallogr.* **39**, 158–168 (2006).
54. Waterman, D. G. *et al.* Diffraction-geometry refinement in the DIALS framework. *Acta Crystallogr. D* **72**, 558–575 (2016).
55. Sauter, N. K. XFEL diffraction: developing processing methods to optimize data quality. *J. Synchrotron Radiat.* **22**, 239–248 (2015).
56. Zhang, Z., Sauter, N. K., van den Bedem, H., Snell, G. & Deacon, A. M. Automated diffraction image analysis and spot searching for high-throughput crystal screening. *J. Appl. Crystallogr.* **39**, 112–119 (2006).
57. Uervirojnangkoorn, M. *et al.* Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. *eLife* **4**, e05421 (2015).
58. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
59. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
60. Padilla, J. E. & Yeates, T. O. A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. *Acta Crystallogr. D* **59**, 1124–1130 (2003).
61. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
62. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
63. Afonine, P. V., *et al.* FEM: feature-enhanced map. *Acta Crystallogr. D* **71**, 646–666 (2015).
64. Cox, N., Pantazis, D. A., Neese, F. & Lubitz, W. Biological water oxidation. *Acc. Chem. Res.* **46**, 1588–1596 (2013).
65. Isobe, H. *et al.* Theoretical illumination of water-inserted structures of the CaMn₄O₅ cluster in the S₂ and S₃ states of oxygen-evolving complex of photosystem II: full geometry optimizations by B3LYP hybrid density functional. *Dalton Trans.* **41**, 13727–13740 (2012).
66. Li, X., Siegbahn, P. E. M. & Ryde, U. Simulation of the isotropic EXAFS spectra for the S₂ and S₃ structures of the oxygen evolving complex in photosystem II. *Proc. Natl Acad. Sci. USA* **112**, 3979–3984 (2015).
67. Ichino, T. & Yoshioka, Y. Theoretical study on mechanism of dioxygen evolution in photosystem II. II. Molecular and electronic structures at the S₃ and S₄ states of oxygen-evolving complex. *Chem. Phys. Lett.* **595**, 237–241 (2014).
68. Hatakeyama, M. *et al.* Structural changes in the S₃ state of the oxygen evolving complex in photosystem II. *Chem. Phys. Lett.* **651**, 243–250 (2016).

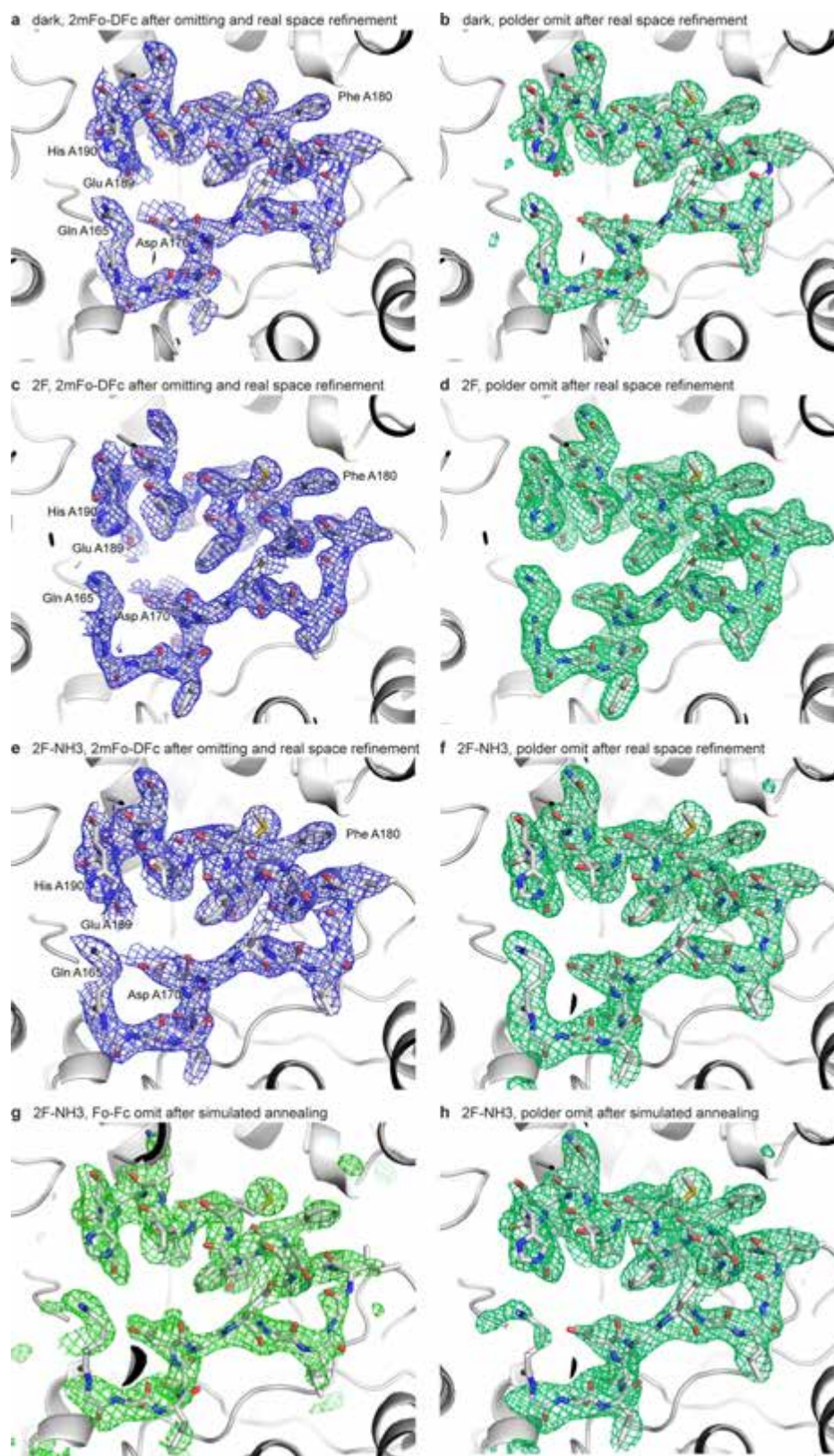
a



Extended Data Figure 1 | See next page for caption.

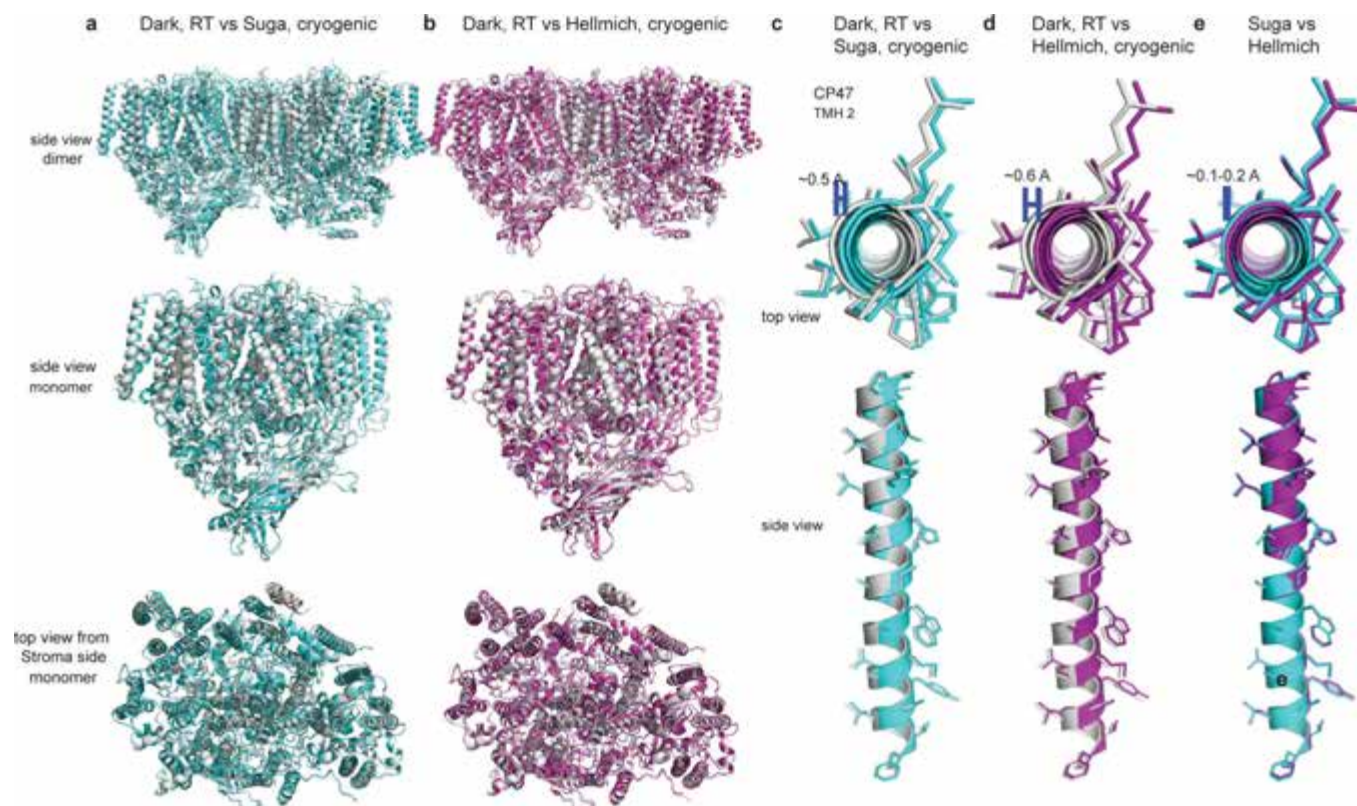
Extended Data Figure 1 | Schematic of the reaction centre and OEC in PS II and packing of the dimeric complex in the crystal lattice. **a**, The reaction centre is shown with cofactors labelled as Pheo (pheophytin), Chl (chlorophyll), PQ (plastoquinone), Q_A , Q_B (primary and secondary acceptor plastoquinones bound to PS II, respectively). The numbering of Mn (purple spheres), oxygen (red/grey spheres) atoms and metal-bound waters in the OEC follows the convention of ref. 21. Upon illumination of PS II, an electron is transferred ~ 35 Å across the membrane from the excited primary electron donor P_{680} to the final electron acceptor Q_B via Chl_{D1} , $Pheo_{D1}$, Q_A , and a non-haem Fe^{II} . After accepting two electrons and undergoing protonation, plastoquinol Q_BH_2 is released from PS II into the membrane matrix. The photo-generated radical cation $P_{680}^{*\cdot+}$ is reduced by a tyrosine residue (Tyr_Z) to generate a neutral tyrosine radical Tyr_Z^\cdot , which acts as an oxidizing agent for water at the OEC. **b**, **c**, **d**, Packing of the dimeric complex observed in the room temperature data for three

different view directions. The unit cell is indicated by a wire frame and axes are labelled. Dimers related by translation are coloured identically. **e–g**, Packing observed in the cryogenic structure in ref. 16 (PDB: 4PJ0) in the same orientations as in **b–d**. **h–j**, Packing observed in the cryogenic XFEL structure in ref. 7 (PDB: 4UB6). The space group is the same in all three cases, but the unit cell dimensions and packing are different. Whereas the *a* and *b* dimensions are very similar between 4PJ0 and the current room temperature data, the *c* axis is elongated. This results in a very similar arrangement of dimers in rows along the *a* axis (compare **d** and **g**), whereas there is a larger spacing between two dimer rows at the cytoplasmic side of the complex (compare black ellipse in **b** and **e**) owing to the elongation of the *c* axis in the room temperature packing. The structure in ref. 7 has a very different arrangement of the dimers and no closely packed rows of dimers are visible (see **c** and **f** compared with **i**, and **d** and **g** compared with **j**).



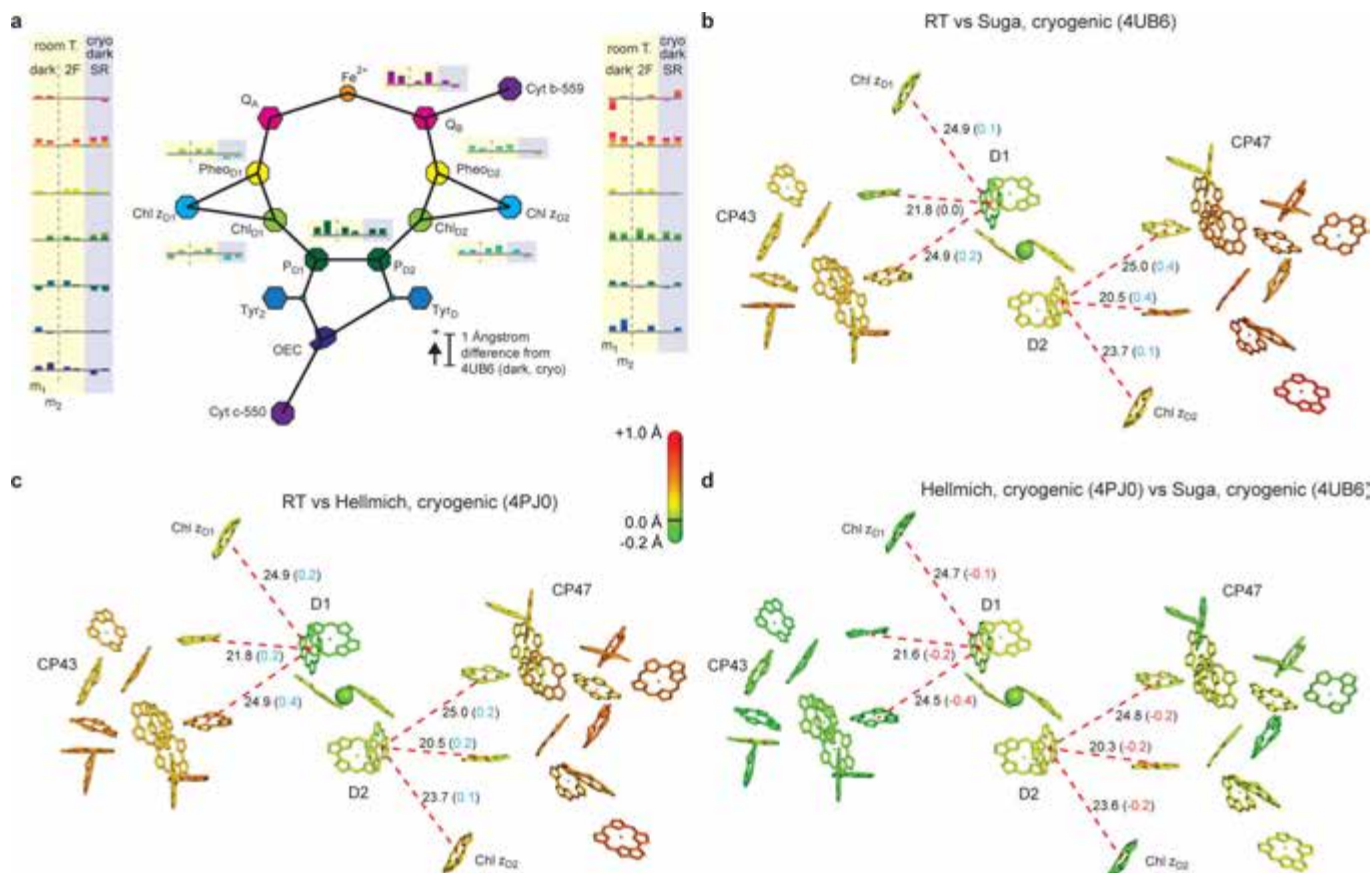
Extended Data Figure 2 | Electron density omit maps of the luminal CD helix and part of the loop region of subunit D1. **a–h**, Obtained from the room temperature dark (**a, b**), 2F (**c, d**), and 2F-NH₃ (**e–h**) datasets. For all maps residues 165–190 of subunit D1 (shown as grey sticks) were omitted followed by three rounds of coordinate and real space refinement of the model with (**g, h**) or without (**a–f**) simulated annealing in *phenix.refine*. **a**, $2mF_o - DF_c$ map (blue, 1.5σ contour) of the dark dataset. **b**, Polder $mF_o - DF_c$ map (green, 4σ contour) of the dark dataset. **c**, $2mF_o - DF_c$

map (blue, 1.5σ contour) of the 2F dataset. **d**, Polder $mF_o - DF_c$ map (green, 4σ contour) of the 2F dataset. **e**, $2mF_o - DF_c$ map (blue, 1.5σ contour) of the 2F-NH₃ dataset. **f**, Polder $mF_o - DF_c$ map (green, 4σ contour) of the 2F-NH₃ dataset. **g**, Standard $mF_o - DF_c$ omit map (green, 3σ contour) of the 2F-NH₃ dataset after simulated annealing. **h**, Polder $mF_o - DF_c$ map (green, 4σ contour) of the 2F-NH₃ dataset after simulated annealing.



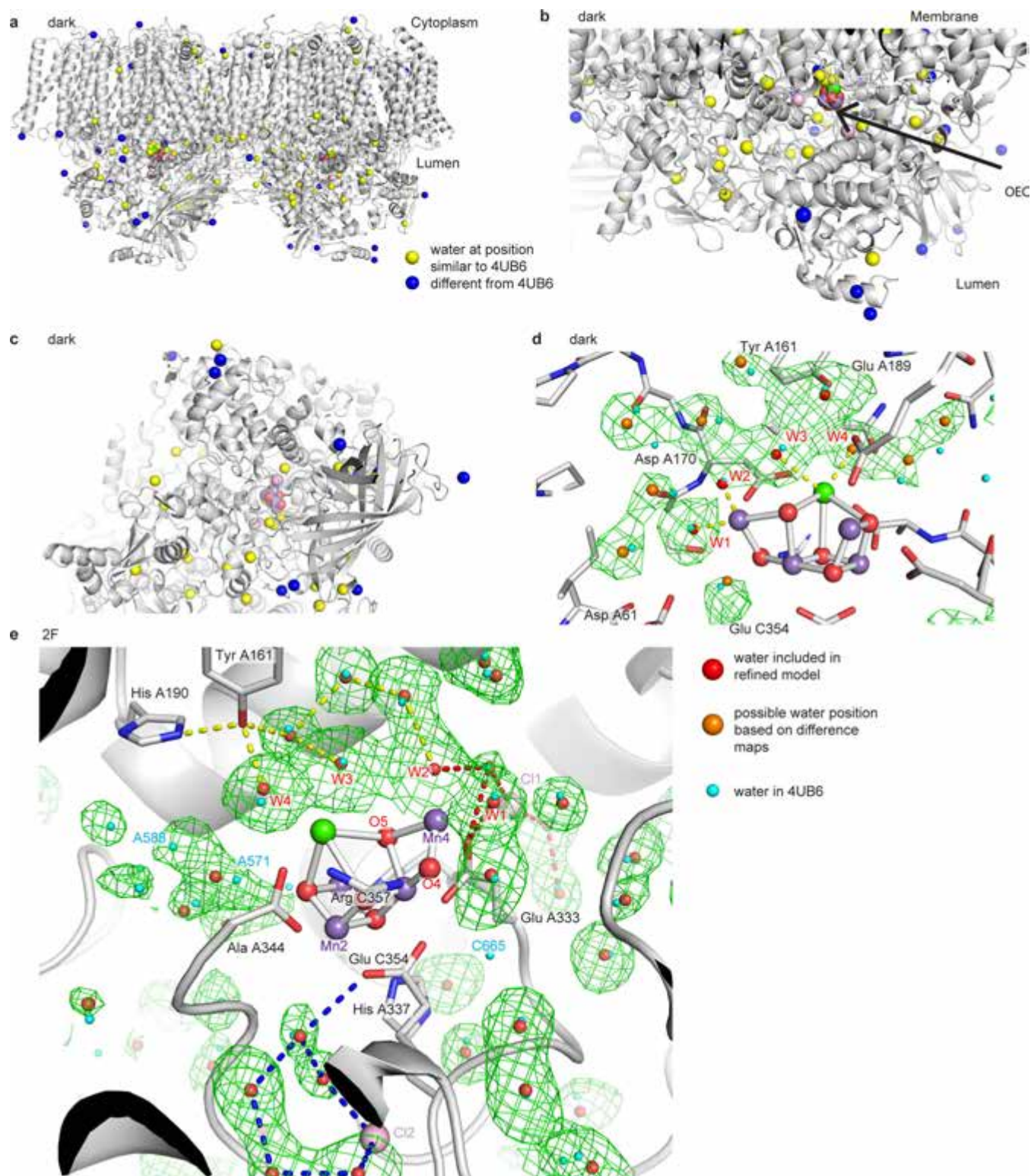
Extended Data Figure 3 | Comparison of dark room temperature structure with cryogenic structures from refs 7 and 16. Our room temperature dark state structure is shown in grey, the Suga XFEL⁷ in cyan and the Hellmich¹⁶ cryogenic structure in purple. **a**, Overlay of the room temperature and Suga⁷ structures. **b**, Overlay of the room temperature and Hellmich¹⁶ structures. A large-scale rigid body motion of the two monomers with respect to each other and an in-plane expansion of each PS II monomer in the room temperature structure are visible. **c–e**, Comparison of TMH 2 of subunit CP47 between the dark state and the cryogenic XFEL structure⁷ (**c**), the dark state and the Hellmich¹⁶

cryogenic structure (**d**) and between the two cryogenic structures (**e**). The two cryogenic datasets reflect crystals with different packing. View is from the cytoplasmic side (top) or along the membrane plane (bottom). Despite the different packing, only a small shift of 0.1–0.2 Å is observed between the two cryogenic structures. In contrast, the room temperature structure exhibits a larger shift of 0.5–0.6 Å in the positions of TMH 2 with respect to the cryogenic structures in both crystal forms, well above the error margin in our data. The cryogenic structures were superposed onto our dark structure model in PyMol (Schrödinger, LLC) using monomer 1 for alignment.



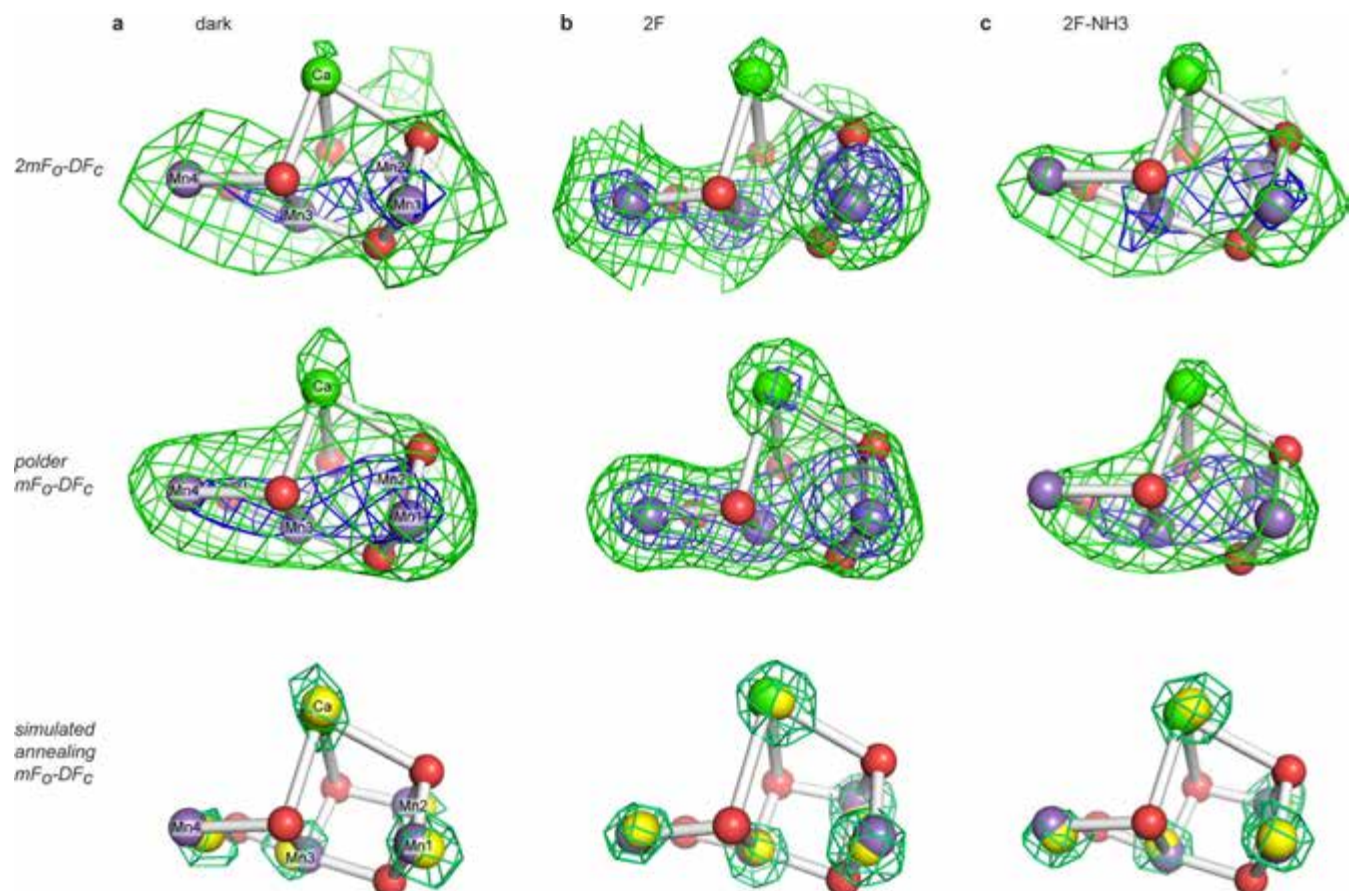
Extended Data Figure 4 | Comparisons of the cofactor-cofactor distances in the crystal structures collected at cryogenic temperature and room temperature. **a**, Distances for central cofactors; the histogram shows the deviation of the cofactor-cofactor distances in the Hellmich cryogenic SR dark state structure (4PJ0¹⁶, highlighted in blue) and in our dark and 2F room temperature structures (highlighted in yellow) from those of the cryogenic XFEL structure reported by Suga *et al.*⁷ (4UB6). Changes are indicated by bars for monomers 1 and 2 (m_1 , m_2), and colour coding of bars matches the colouring of the associated pair of cofactors in the diagram. Among the differences, there is a consistent elongation in the distances involving Chl_L as well as in the Q_B - $Cyt\ b-559$ and OEC - $Cyt\ c-550$ distances of both monomers in the room temperature data. In other cases, expansion of individual cofactor distances is observed in both room

temperature structures and 4PJ0 relative to 4UB6 (for example, P_{D2} - Chl_{D2} , Q_B - $Pheo_{D2}$), and in the case of P_{D1} - P_{D2} on average the elongation is more pronounced at room temperature than in the cryogenic structures. Changes in Chl positions between the room temperature dark structure and 4UB6 (**b**), between the room temperature dark structure and 4PJ0, which have the same dimer-dimer packing (**c**), and between 4PJ0 and 4UB6 (**d**). The distances of the Chl ring centres from the membrane normal passing through the centre between $Pheo_{D1}$ and $Pheo_{D2}$ are computed and relative changes with respect to the values obtained from 4UB6 or 4PJ0 are shown as colour coding on a rainbow scale from green (0.2 Å contraction) to red (1.0 Å expansion). For selected Chl - Chl pairs, distances are given with the absolute change in parentheses.



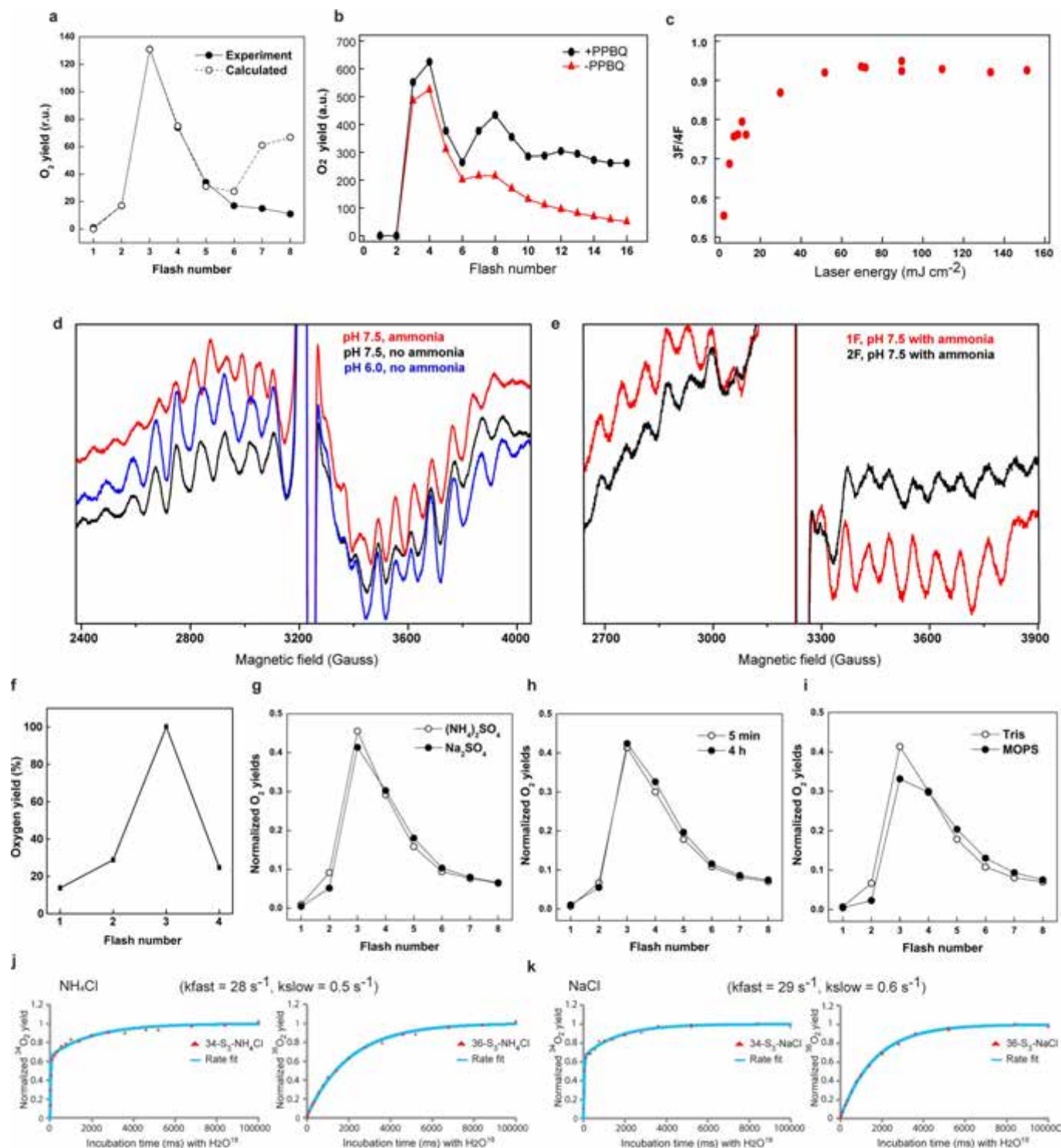
Extended Data Figure 5 | Location of water molecules observed in the room temperature structure. **a**, Water positions refined in the dark room temperature structure of the PS II dimer are indicated by blue and yellow spheres. View is along the membrane plane with the cytoplasm on top and lumen on the bottom. Waters whose positions coincide with waters located in the cryogenic XFEL structure⁷ (4UB6) are coloured in yellow and waters in other positions are coloured in dark blue. **b**, Enlarged view of the lumenal region showing the OEC of one monomer (magenta, red and green spheres for Mn, O and Ca) as well as the two Cl⁻ (pink spheres) located close to the OEC. **c**, Waters located at the luminal side of one monomer. View is from the luminal side onto the membrane plane, with colour coding as in **a** and **b**. **d**, Polder omit maps (2.5 σ contour, green

mesh) for waters in the direct vicinity of the OEC in the dark state. Waters included in the refined model are indicated as red spheres, additional waters placed based on polder maps as orange spheres, and waters from 4UB6 are shown in light cyan. **e**, Possible water networks next to the OEC. Waters included in the refined model of the 2F state are indicated as red spheres and waters from 4UB6 are shown in light cyan. Polder omit maps (2.0 σ contour, green mesh, carved at 2 Å around water positions from 4UB6) confirm the positions of the refined waters and indicate the presence of additional waters (for example, A571, A588), but no omit map density was observed at the position of water C665. The starting points of three water/proton channels postulated in ref. 21 are indicated by dashed red, yellow, and blue lines.



Extended Data Figure 6 | Room temperature electron density of the Mn_4CaO_5 cluster. **a**, The $2mF_o - DF_c$ electron density (top) contoured at 4.0σ (green) and 8.0σ (blue mesh) and the polder $mF_o - DF_c$ electron density (middle) of the dark dataset after omitting the OEC and real space refinement contoured at 8.0σ (green) and 14.0σ (blue mesh). At the bottom, the $mF_o - DF_c$ electron density after omitting individual metal atoms and refining with simulated annealing is shown contoured at 4.0σ (Ca), 7.0σ (Mn1, Mn2), 8.0σ (Mn3) and 4.0σ (Mn4). The model of the OEC is shown with Mn as magenta, Ca as green and oxygen as red spheres overlaid with yellow spheres indicating the centres of the obtained omit densities. **b**, The $2mF_o - DF_c$ electron density (top) contoured at 3.0σ (green) and 6.0σ (blue mesh) and the polder $mF_o - DF_c$ electron

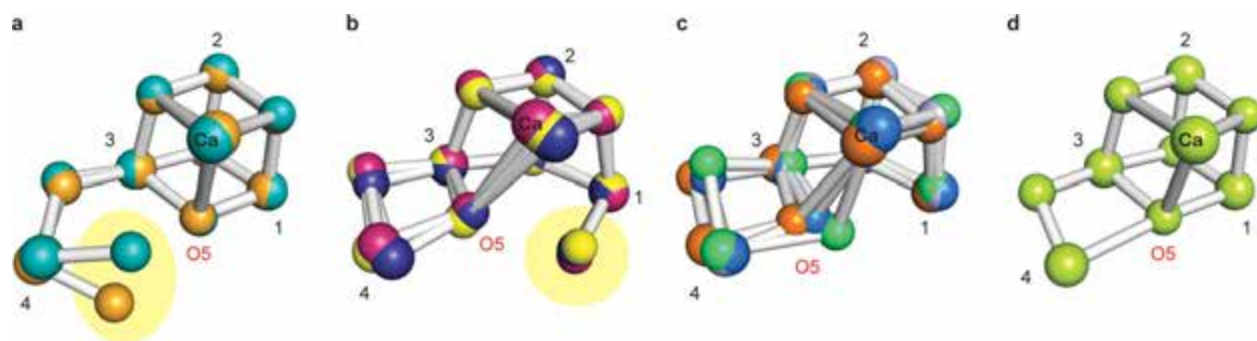
density (middle) of the 2F dataset after omitting the OEC and real space refinement contoured at 8σ (green) and 14σ (blue mesh). At the bottom, the $mF_o - DF_c$ electron density after omitting individual metal atoms and refining with simulated annealing is shown contoured at 12.0σ (Ca, Mn1, Mn3), 13.0σ (Mn2) and 10.0σ (Mn4), with colour coding as in **a**. **c**, The $2mF_o - DF_c$ electron density (top) contoured at 5.0σ (green) and 8.0σ (blue mesh) and the polder $mF_o - DF_c$ electron density (middle) of the 2F-NH₃ dataset after omitting the OEC and real space refinement contoured at 11σ (green) and 16σ (blue mesh). At the bottom, the $mF_o - DF_c$ electron density after omitting individual metal atoms and refining with simulated annealing is shown contoured at 5.0σ (Ca), 10.0σ (Mn1), 8.0σ (Mn2, Mn4) and 11.0σ (Mn3), with colour coding as in **a**.



Extended Data Figure 7 | See next page for caption.

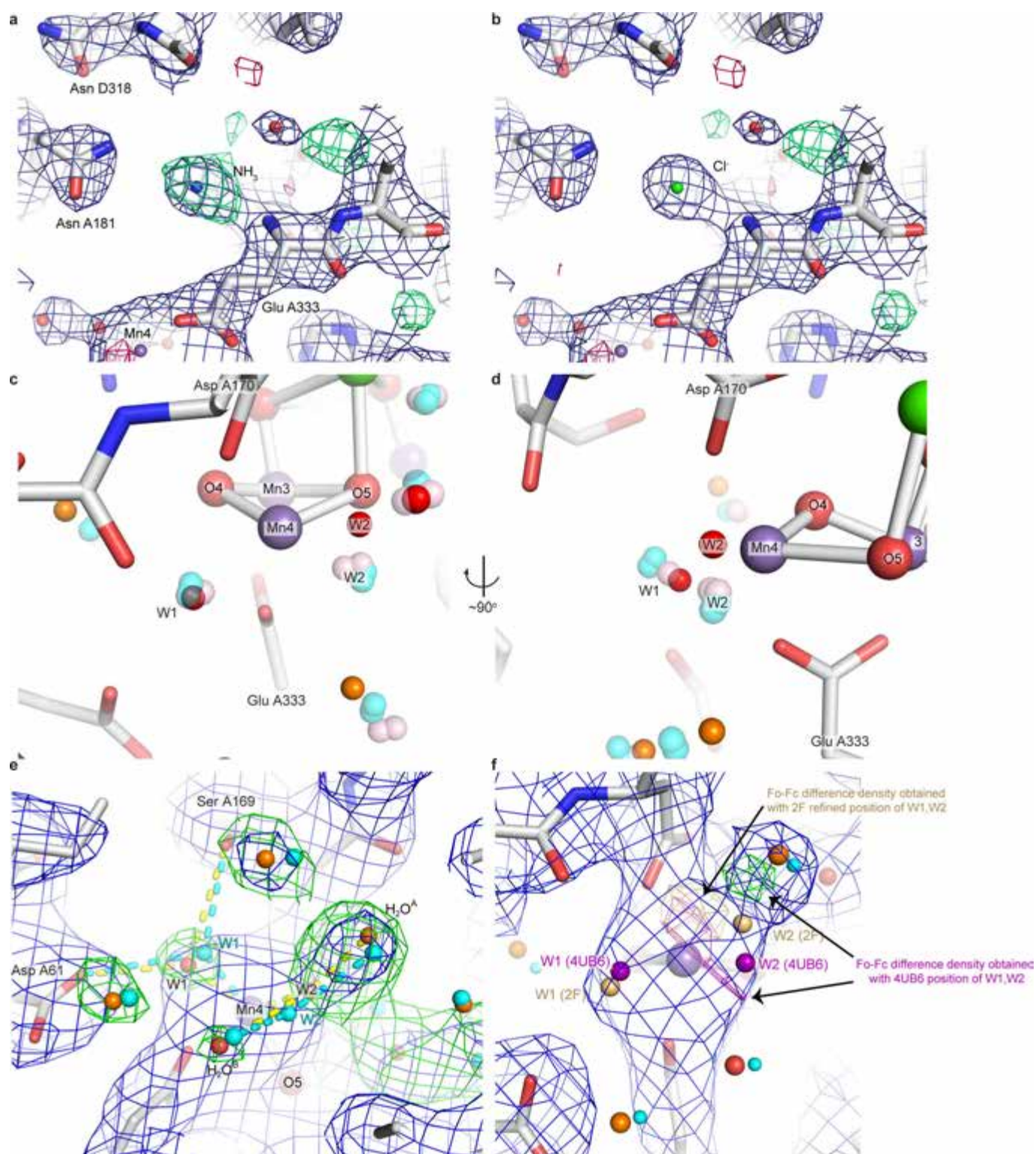
Extended Data Figure 7 | Characterization of PS II samples and substrate water exchange rates in the S_3 state. **a**, Flash-induced O_2 evolution pattern (FIOP) of a suspension of the native PS II core complexes (PSIIcc) at pH 6.5. Fit parameter: 100% S_1 in the dark, miss 20%, double hit 4%, damping 2%, fit done on first five flashes. In the 2F sample, if double hits are excluded, $\sim 60\%$ of the S_3 population was calculated, with $\sim 30\%$ of S_2 and $\sim 4\%$ of S_1 . **b**, FIOP of a suspension of the PS IIcc (TRIS, pH 7.5, 100 mM $(NH_4)_2SO_4$) with and without PPBQ measured with 12 s between flashes using the thin layer MIMS setup (see Methods) and 532 nm laser flash illumination: an increase in O_2 yield by 15% was observed if PPBQ was added. **c**, Light saturation in the thin layer MIMS set up resembling the illumination conditions of the DOT approach. The ratio of the oxygen yield of the third flash over that of the fourth flash is plotted as a qualitative measure for the miss parameter, which is minimal when the ratio is large. **d**, EPR spectra of native (pH 6.0 and 7.5) and ammonia-treated (pH 7.5) PS II solutions after continuous illumination at 195 K for 1 min followed by annealing to 260 K for 30 s. Spectrometer condition: microwave frequency, 9.23 GHz; field modulation amplitude, 32 G at 100 KHz; microwave power, 20 mW. The spectra were collected at 7 K. **e**, EPR spectra of ammonia-treated (pH 7.5) PS II solution after applying one (red) or two (black) flashes. Spectrometer conditions are as in **a**. **f**, O_2 -flash pattern of PS II crystals at pH 7.5 (TRIS, $(NH_4)_2SO_4$) measured by MIMS with the replica set up for jet illumination described above. **g**, FIOP (Joliot-type electrode) of PSIIcc at pH 7.5 and 20 °C in TRIS buffer with either 100 mM $(NH_4)_2SO_4$ or 100 mM Na_2SO_4 addition.

The O_2 yields for each sample were normalized to the O_2 yields induced by flashes 3–6 (Y3–6). No artificial electron acceptors were added. The flash frequency of the Xe-flash lamp was 2 Hz, and the Chl concentration 0.4 mM. Data are the average of three technical replicates. From the data, a miss parameter of 23–25% and an S_3 state population of 50–53% can be extracted for both sample types. Double hits, caused by the Xe flash lamp, are 3–6%, and are absent under laser flash illumination used during the XFEL experiments. The total O_2 yield of the ammonia-containing sample was 66% of the Na_2SO_4 control, and 52% of the FIOP at pH 6.5 (**a**). **h**, FIOP of PS II core sample incubated in TRIS and 100 mM $(NH_4)_2SO_4$ for 5 min at pH 7.6 versus one that was incubated for 4 h at room temperature. No degradation of the sample was observed over time (O_2 yields of both FIOPs normalized to Y3–6 of 4-min trace). **i**, FIOP of PS II core sample containing TRIS and 100 mM $(NH_4)_2SO_4$ versus one containing MOPS and 100 mM $(NH_4)_2SO_4$. Both FIOPs normalized to Y3–6 of TRIS containing sample. **j**, **k**, Substrate water exchange was measured for the S_3 state of PS II core complexes at pH 7.6/20 °C as described^{2,4,12} in HEPES buffer containing 100 mM NH_4Cl (**j**) or 100 mM $NaCl$ (**k**). The left panels show the biphasic rise of the mass 34 peak ($^{16}O^{18}O$), while the right side shows the simultaneously recorded monophasic rise of the 36 peak of the double exchanged $^{18}O^{18}O$. Red symbols represent the individually measured data points, while the blue lines are the kinetic simulations. Nearly identical rates for the exchange of the fast (k_{fast}) and slow (k_{slow}) substrates were found with and without ammonia.



Extended Data Figure 8 | Possible Mn_4CaO_5 complex models for the S_3 -state proposed in the literature. The models are grouped into four classes (**a–d**, see below). Mn are numbered (1–4) as in the main text. **a**, Models with an inserted water (highlighted in yellow) on the left side (closed cubane proposed as a transient S_3 -state in ref. 64 in teal; ref. 65 in light orange) shift Mn4 the furthest out. **b**, Models with an inserted water (highlighted in yellow) on the right side (model in ref. 64 in yellow; ref. 66 in pink; ref. 67 in dark blue) closely resemble models with an open cubane, and were proposed initially by Li and Siegbahn²³. **c**, Other models

with no inserted water: from ref. 67 in marine blue; 4UB6⁷ in green; from ref. 67 in lavender; from ref. 65 in dark orange). **d**, Only one proposed model featured a closed cubane with no inserted water (ref. 68 in yellow-green). Note that, except for the type **a** structure (a complete cubane plus mono- μ -oxo bridged Mn4) and the type **d** structure (a closed cubane with no water inserted), the Mn atomic positions are very similar in all models within 0.26 Å. Even between the type **a** and other models, the Mn4 positions differ only by 0.73 Å. On the other hand, the O5 position is expected to differ among the models by at most 1.52 Å.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | The electron density of the Cl^- binding sites and environment of the W1 and W2 sites at the OEC in the 2F-NH₃ samples in the 2F dataset. **a**, The Cl^- binding site 1 with Cl^- (green sphere) at its refined position. The $2mF_o - DF_c$ map (blue mesh) is shown at 1.5σ , and the $mF_o - DF_c$ map (green/red) is shown at $\pm 3\sigma$. **b**, Cl^- binding site 1 with ammonia (blue sphere) instead of Cl^- included in the model. The lack of difference density at the Cl^- position in the refined model and the positive difference density observed when ammonia is substituted for Cl^- indicate that ammonia does not account for the electron density and that Cl^- is a good model for the observed density. **c, d**, Comparison of the positions of W1 and W2 among the 2F-NH₃ (red), 2F (light pink) and cryogenic S₁-XFEL⁷ (light cyan) structures in two different orientations. The outcome of two different alignment procedures (to optimize overlap of either OEC Mn atoms or the surrounding protein ligands) are shown, illustrating the error in these alignments. A small shift of both W1 and W2 upon transition from the cryogenic S₁-XFEL to the 2F structure is visible. In the 2F-NH₃ model, W1 is shifted slightly

further along the same direction as the dark-2F difference. In contrast, the displacement of W2 in the 2F-NH₃ model is significantly larger than that between the S₁-XFEL and 2F structures. **e**, $2mF_o - DF_c$ electron density (blue, 1.0σ) of the 2F-NH₃ dataset around Mn4, and polder omit maps at 3.5σ (green). Red spheres, water refined in the current model; orange spheres, water placed in the polder maps but not included in the refined model; cyan spheres, water positions from the cryogenic XFEL structure. **f**, When calculating the $mF_o - DF_c$ electron density using the refined positions of W1 and W2 (light orange spheres) from the 2F-NH₃ data, only a negative peak is observed (orange mesh, -3σ contour) while using W1 and W2 positions (purple spheres) from the S₁-XFEL structure yields clear positive and negative peaks (green and purple mesh, $\pm 3\sigma$ contour, colouring of other waters as in **e**). This indicates that the W2 position from the S₁-XFEL structure does not provide a good fit to the observed electron density. However, at the present resolution the observed difference densities may be influenced by other effects, for example, Fourier series truncations and the strong density of H₂O^A close to W2.

Extended Data Table 1 | Data collection and refinement statistics for the room temperature dark (S₁), 2F and 2F-NH₃ data

	S ₁ dataset (5KAF)	2F dataset (5TIS)	2F-NH ₃ dataset (5KAI)
Data collection			
Space group	<i>P</i> 2 ₁ 2 ₁ 2 ₁	<i>P</i> 2 ₁ 2 ₁ 2 ₁	<i>P</i> 2 ₁ 2 ₁ 2 ₁
Cell dimensions <i>a</i> , <i>b</i> , <i>c</i> (Å)	117.7±1.2, 223.8±2.5, 330.8±2.6	117.9, 223.1, 310.7 ^a	117.9±1.6, 224.3±3.3, 331.0±2.9
α , β , γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Wavelength (Å)	1.7493	1.3010	1.7492
Resolution (Å)	43.13-3.00 (3.05-3.00) ^b	44.28-2.25 (2.29-2.25)	43.58-2.80 (2.85-2.80)
Unique reflections	170444 (7698)	385065 (19084)	213400 (10235)
Completeness (%)	97.30 (88.80)	99.98 (99.98)	99.00 (95.96)
Multiplicity	8.48 (2.91)	158.47 (10.36)	13.70 (4.17)
<i>I</i> / σ (<i>I</i>)	10.51 (2.26)	19.83 (2.04)	11.62 (1.94)
<i>CC</i> _{1/2}	53.2 (13.2)	97.4 (1.8)	54.2 (7.3)
Wilson B-factor	61.05	42.98	60.41
Collection instrument	LCLS endstation CXI	LCLS endstations XPP and MFX	LCLS endstation CXI
Refinement			
<i>R</i> _{work} (%)	26.37 (33.50)	19.49 (34.04)	24.97 (33.98)
<i>R</i> _{free} (%)	30.30 (37.98)	23.08 (38.09)	29.97 (33.98)
No. atoms	50162	51757	50284
Protein residues	5319	5319	5316
Ligands	174	189	180
Waters	124	1179	107
Average <i>B</i> factor	47.3	45.9	56.6
R.m.s. deviations			
Bond lengths (Å)	0.004	0.005	0.004
Bond angles (°)	0.479	0.502	0.592
Ramachandran favored (%)	96.18	97.20	95.20
Ramachandran outliers (%)	0.19	0.29	0.57
MolProbity clashscore	4.45	4.45	6.55

^aUnit cell parameters were constrained for this dataset.^bValues in parentheses are for the highest-resolution shell.

Structure of CC chemokine receptor 2 with orthosteric and allosteric antagonists

Yi Zheng¹, Ling Qin¹, Natalia V. Ortiz Zacarías², Henk de Vries², Gye Won Han³, Martin Gustavsson¹, Marta Dabros⁴, Chunxia Zhao¹, Robert J. Cherney⁴, Percy Carter⁴, Dean Stamos⁵, Ruben Abagyan¹, Vadim Cherezov³, Raymond C. Stevens⁶, Adriaan P. IJzerman², Laura H. Heitman², Andrew Tebben⁴, Irina Kufareva¹ & Tracy M. Handel¹

CC chemokine receptor 2 (CCR2) is one of 19 members of the chemokine receptor subfamily of human class A G-protein-coupled receptors. CCR2 is expressed on monocytes, immature dendritic cells, and T-cell subpopulations, and mediates their migration towards endogenous CC chemokine ligands such as CCL2 (ref. 1). CCR2 and its ligands are implicated in numerous inflammatory and neurodegenerative diseases² including atherosclerosis, multiple sclerosis, asthma, neuropathic pain, and diabetic nephropathy, as well as cancer³. These disease associations have motivated numerous preclinical studies and clinical trials⁴ (see <http://www.clinicaltrials.gov>) in search of therapies that target the CCR2–chemokine axis. To aid drug discovery efforts⁵, here we solve a structure of CCR2 in a ternary complex with an orthosteric (BMS-681 (ref. 6)) and allosteric (CCR2-RA-[R])⁷ antagonist. BMS-681 inhibits chemokine binding by occupying the orthosteric pocket of the receptor in a previously unseen binding mode. CCR2-RA-[R] binds in a novel, highly druggable pocket that is the most intracellular allosteric site observed in class A G-protein-coupled receptors so far; this site spatially overlaps the G-protein-binding site in homologous receptors. CCR2-RA-[R] inhibits CCR2 non-competitively by blocking activation-associated conformational changes and formation of the G-protein-binding interface. The conformational signature of the conserved microswitch residues observed in double-antagonist-bound CCR2 resembles the most inactive G-protein-coupled receptor structures solved so far. Like other protein–protein interactions, receptor–chemokine complexes are considered challenging therapeutic targets for small molecules, and the present structure suggests diverse pocket epitopes that can be exploited to overcome obstacles in drug design.

A ternary complex between an engineered construct of human CCR2 isoform B (further referred to as CCR2-T4L or simply CCR2), an orthosteric antagonist BMS-681 (compound 13d in ref. 6), and an allosteric antagonist CCR2-RA-[R]⁷ was crystallized using the lipidic cubic phase (LCP) method⁸, and the structure was determined to 2.8 Å resolution (Extended Data Table 1 and Extended Data Fig. 1). Simultaneous addition of two compounds markedly stabilized detergent-solubilized CCR2-T4L compared with twice the concentration of each compound individually (Fig. 1a), suggesting concurrent binding of CCR2-RA-[R] and BMS-681 to the receptor. The presence of both compounds was critical for crystallization.

In the structure, CCR2 adopts the canonical fold of class A G-protein-coupled receptors (GPCRs) with seven transmembrane (TM) helices connected by three extracellular (EC) and three intracellular (IC) loops (Fig. 1b). Both compounds are visible in the electron density (Fig. 1b–d); BMS-681 binds in the extracellular orthosteric pocket (Fig. 1b, c) while CCR2-RA-[R] is located more than 30 Å away (Fig. 1b, d), in a site that is the most intracellular allosteric pocket observed in class A GPCRs so far (Fig. 1e). The binding site of CCR2-RA-[R] spatially

overlaps with the G-protein-binding site in homologous receptors (Fig. 1f). As for other chemokine receptors^{9–12}, CCR2 is expected to have two conserved disulfide bonds in its extracellular domains, with Cys32–Cys277 connecting the amino (N) terminus to ECL3 (NT–ECL3), and Cys113–Cys190 connecting TM3 to ECL2. Electron density is apparent for the ECL2–TM3 disulfide bond but not for the N-terminal residues 1–36 or the NT–ECL3 disulfide bond (Fig. 1b, c). Because the NT–ECL3 disulfide bond has been shown to be important for CCR2 signalling¹³, its absence is unlikely to be an inherent feature of the receptor; instead, it might be caused by strain of the bond in the ligand-bound state of the receptor¹⁴, possibly exacerbated by solvent exposure and radiation damage of the crystals¹⁵.

As with other chemokine receptors, the extracellular orthosteric pocket of CCR2 can be divided into a major and a minor subpocket, defined by helices III–VII, and helices I–III and VII, respectively, and separated by residues Y120^{3,32} and E291^{7,39} (superscript indicates residue number according to Ballesteros–Weinstein nomenclature). BMS-681 binds predominantly in the minor subpocket (Fig. 2a, b) and buries 366.3 Å² of surface area. The 6-trifluoromethyl quinazoline moiety protrudes between helices I and VII towards the lipid bilayer, while the tri-substituted cyclohexane packs against W98^{2,60}. The γ -lactam secondary exocyclic amine forms a hydrogen bond with the hydroxyl of T292^{7,40}, which is critical for binding of chemically related compounds such as BMS-558 (compound 22 in ref. 16) and the Teijin lead series^{17,18}. This amine is also within hydrogen-bonding distance from the backbone carbonyl of Q288^{7,36}. The carbonyl oxygen of the γ -lactam forms a hydrogen bond with Y49^{1,39}, which itself is hydrogen-bonded to the side chain of T292^{7,40}. The N1 nitrogen of the quinazoline is within 4 Å of the Q288^{7,36} side chain. The protonated tertiary amine on the cyclohexane ring is proximal to a structured water molecule in the binding site. Some CCR2 antagonists, particularly those containing a basic amine, are known to depend on the conserved E291^{7,39} in the receptor¹⁹; however, no direct interaction is observed between E291^{7,39} and BMS-681. The receptor-bound, bioactive conformation of BMS-681 is strikingly similar to the crystallographic conformation of free BMS-681 (Fig. 2c and Extended Data Table 2), suggesting the absence of internal strain in the bound state.

BMS-681 engages several residues that are critical for CCL2 binding and/or activation of CCR2 (refs 17, 18) including Y49^{1,39}, W98^{2,60}, Y120^{3,32}, and T292^{7,40}. Thus, it seems to directly compete with chemokine binding to the orthosteric pocket. Additionally, by inserting between helices I and VII, BMS-681 may put strain onto residues C32–V37 connecting TM1 to ECL3, destabilize the conserved NT–ECL3 disulfide bond (absent in the structure), and prevent the N terminus and TM1 from adopting a productive chemokine binding conformation observed in homologous receptor–chemokine structures^{11,12} (Extended Data Fig. 2).

¹Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ²Division of Medicinal Chemistry, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Leiden 2333 CC, The Netherlands. ³Bridge Institute, Departments of Chemistry and Physics & Astronomy, University of Southern California, Los Angeles, California 90089, USA. ⁴Bristol-Myers Squibb Company, Princeton, New Jersey 08543, USA. ⁵Vertex Pharmaceuticals Inc., 11010 Torreyana Road, San Diego, California 92121, USA. ⁶The Bridge Institute, Departments of Biological Sciences and Chemistry, University of Southern California, Los Angeles, California 90089, USA.

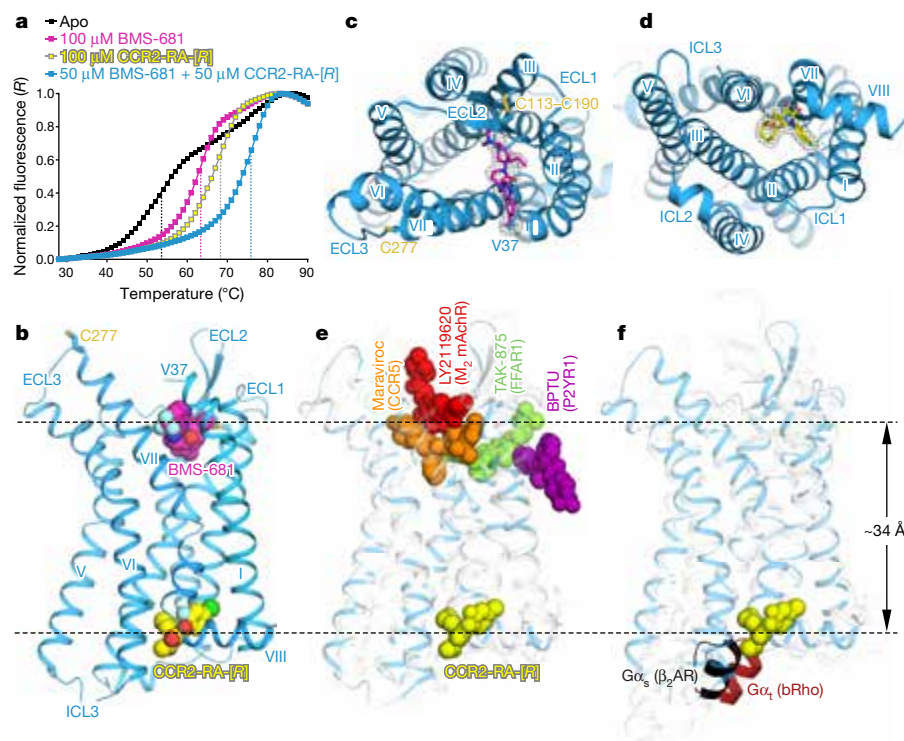


Figure 1 | Structure of a complex between CCR2, BMS-681 and CCR2-RA-[R] and comparison with other allosteric modulators of class A GPCRs. **a**, Thermal denaturation curves demonstrate higher stability of CCR2-T4L in the presence of both BMS-681 and CCR2-RA-[R] compared with each compound individually. Data are representative of three independent experiments conducted on different days. **b**, Overall view of double-antagonist-bound CCR2. **c**, **d**, Structure viewed from the extracellular (**c**) and intracellular (**d**) side with simulated annealing omit maps of BMS-681 (**c**) and CCR2-RA-[R] (**d**) shown at 3σ . **e**, CCR2-RA-[R] compared with other allosteric ligands crystallized with class A GPCRs (Protein Data Bank (PDB) accession numbers 4MBS, 4XNV, 4PHU, and 4MQT). **f**, CCR2-RA-[R] compared with the carboxy (C)-terminal helix of $G\alpha_s$ bound to the β_2 adrenergic receptor and transducin peptide bound to rhodopsin (PDB accession numbers 3SN6 and 4X1H).

On the opposite side of the receptor, CCR2-RA-[R] is caged by the intracellular ends of helices I–III and VI–VIII and buries 297.8 \AA^2 of surface area. The inner hydrophobic part of the cage is made by V63^{1,53}, L67^{1,57}, L81^{2,43}, L134^{3,46}, A241^{6,33}, V244^{6,36}, I245^{6,37}, Y305^{7,53}, and F312^{8,50}, while the outer (cytosol-facing) polar part consists of T77^{2,39}, R138^{3,50}, G309^{8,47}, K311^{8,49}, and Y315^{8,53} (Fig. 2d, e), as well as the backbones of engineered R237^{6,29} and K240^{6,32}. The binding pocket of CCR2-RA-[R] is highly enclosed and possesses a balanced combination of hydrophobic and polar features, all of which favours pocket ‘druggability’²⁵. Owing to the lack of a side-chain on G309^{8,47}, the hydroxyl and pyrrolone carbonyl groups of CCR2-RA-[R] can hydrogen-bond to the exposed backbone amides of E310^{8,48}, K311^{8,49}, and F312^{8,50} (Fig. 2d, e). The acetyl group of the compound resides near the terminal amine of K311^{8,49}. The critical roles of V244^{6,36}, Y305^{7,53}, K311^{8,49}, and F312^{8,50} in CCR2-RA-[R] binding were established by an earlier mutagenesis study²⁰. Because homologues of several residues in the CCR2-RA-[R] binding pocket directly couple to the G protein in bovine rhodopsin²¹ and the β_2 adrenergic receptor ($\beta_2\text{AR}$)²² structures (Extended Data Fig. 3), CCR2-RA-[R] appears to sterically interfere with G-protein binding to CCR2.

The structure suggests an interesting symmetrical mechanism for the concurrent antagonistic action of the two compounds. BMS-681 interferes with chemokine binding directly and with G-protein coupling indirectly, by stabilizing an inactive, presumably G-protein-incompatible⁶, conformation of the receptor. Conversely, CCR2-RA-[R] directly prevents G-protein coupling and allosterically inhibits binding of the CCL2 chemokine²³, which, like most GPCR agonists, requires an active, G-protein-associated receptor for high affinity binding²³. Bi-directional allosteric communication between the extra- and intracellular sides of the receptor is reminiscent of that previously observed in adenosine A_{2A} receptor ($AA_{2A}R$)²⁴ and $\beta_2\text{AR}$ ²⁵ using allosteric inverse agonist antibodies/nanobodies that target the same epitope as CCR2-RA-[R]. Similar to these antibodies, CCR2-RA-[R] was previously shown to allosterically enhance, and to be allosterically enhanced by, binding of orthosteric antagonists²³, demonstrating positive binding cooperativity.

We further characterized this cooperativity by studying the binding of BMS-681 to wild-type CCR2 and the crystallization construct CCR2-T4L using previously characterized radioactive probes [^3H]INCB-3344

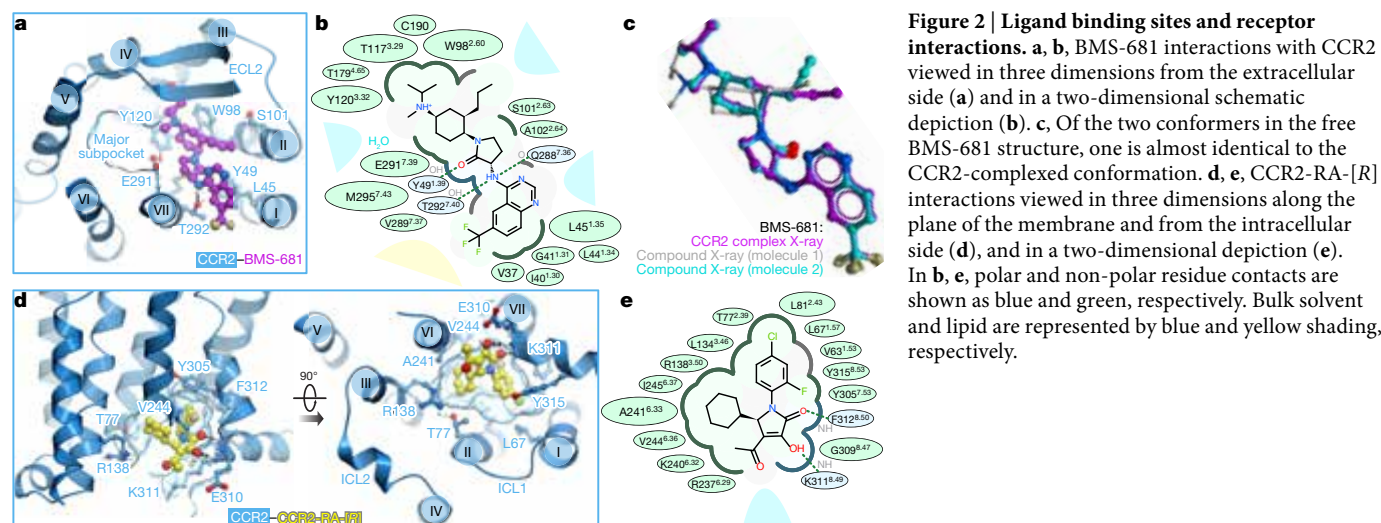


Figure 2 | Ligand binding sites and receptor interactions. **a**, **b**, BMS-681 interactions with CCR2 viewed in three dimensions from the extracellular side (**a**) and in a two-dimensional schematic depiction (**b**). **c**, **d**, Of the two conformers in the free BMS-681 structure, one is almost identical to the CCR2-complexed conformation. **d**, **e**, CCR2-RA-[R] interactions viewed in three dimensions along the plane of the membrane and from the intracellular side (**d**), and in a two-dimensional depiction (**e**). In **b**, **e**, polar and non-polar residue contacts are shown as blue and green, respectively. Bulk solvent and lipid are represented by blue and yellow shading, respectively.

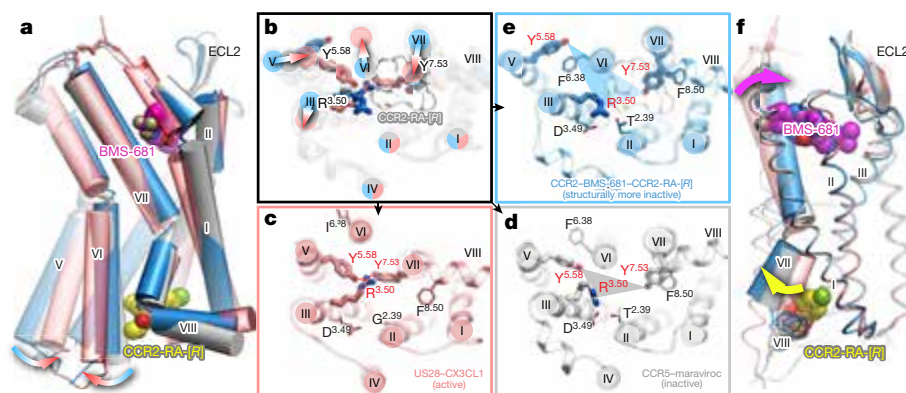


Figure 3 | Crystallographic conformation of double-antagonist-bound CCR2 has pronounced structural signatures of an inactive state. Structures of active (US28, salmon), inactive (CCR5, grey), and more inactive (CCR2, blue) chemokine receptors viewed along the plane of the membrane (a) and across the membrane from the intracellular side (b–e). b, Overlay of structures; arrows show the direction of activation-associated conformational changes; sticks show conserved Y^{5.58}, R^{3.50} and Y^{7.53}; the white mesh is CCR2-RA-[R]. c–e, Detailed, single-receptor depictions of b. f, Although located ~30 Å apart, the orthosteric (BMS-681, magenta) and the allosteric (CCR2-RA-[R], yellow) ligands cooperate in stabilizing an inactive conformation of CCR2 through helix VII.

(ref. 23) (orthosteric) and [³H]CCR2-RA (allosteric). In equilibrium competition binding assays on wild-type CCR2, both INCB-3344 and CCR2-RA-[R] displaced their homologous radioligand with half-maximum inhibitory concentration (IC₅₀) values of 17 and 13 nM, respectively (Extended Data Fig. 4a, b and Extended Data Table 3), comparable to previously reported values²³. Compared to wild-type CCR2, the affinity of both antagonists towards CCR2-T4L was improved by approximately twofold, suggesting a slight engineering-related shift towards the inactive state. BMS-681 fully displaced [³H]INCB-3344 with nanomolar affinities for both constructs, but did not displace [³H]CCR2-RA. Instead, at 1 μM concentration it enhanced the binding of [³H]CCR2-RA by >30% (Extended Data Fig. 4a, b and Extended Data Table 3).

In kinetic radioligand experiments, the presence of BMS-681 also increased total binding of [³H]CCR2-RA to both wild-type CCR2 and CCR2-T4L, with the increase as high as 62% in the case of CCR2-T4L (Extended Data Fig. 4c, d and Extended Data Table 4). BMS-681 (1 μM) decreased the dissociation rate constant of [³H]CCR2-RA, while producing a slight increase (wild-type CCR2) or no change (CCR2-T4L) in the observed association rate constants. Moreover, for CCR2-T4L, the presence of BMS-681 changed the biphasic dissociation profile of [³H]CCR2-RA to monophasic, suggesting stabilization of the receptor population in a homogenous conformational state (Extended Data Table 4). Along with the stability and equilibrium binding data, these results further corroborate the hypothesis that BMS-681 and CCR2-RA-[R] cooperatively stabilize a preferred inactive conformation of CCR2-T4L.

We next analysed the structure of double-antagonist-bound CCR2-T4L to better understand this conformation. The plethora of existing class A GPCR structures suggests a conserved conformational signature of an active receptor state²⁶. This signature involves increased separation between the intracellular end of helix VI and the rest of the TM bundle, an inward repositioning and rotation of helix VII, and concerted repacking of the highly conserved microswitches R^{3.50} (of the DR^{3.50}Y motif), Y^{5.58}, and Y^{7.53} (of the NPxxY^{7.53} motif) (Fig. 3a, b) to

form an intracellular binding interface for G protein. Furthermore, rather than adopting either an 'on' or 'off' state, receptors can occupy an ensemble of intermediate conformations²⁷. The active state signature is fully represented in US28, the only agonist-bound chemokine receptor crystallized so far¹² (Fig. 3a–c). By contrast, the double-antagonist-bound CCR2 structure appears to occupy the opposite end of the activation spectrum as it shares the conformational microswitch signatures of the most inactive GPCR structures observed thus far (Fig. 3a–e).

As in the inactive CCR5–maraviroc complex¹⁰, the intracellular ends of CCR2 helices III and VI are close together, and the conserved R^{3.50} interacts with D^{3.49} and T^{2.39}, effectively disrupting the G-protein-binding pocket (Fig. 3b, d, e). Similarly, in both CCR2 and CCR5 structures, the intracellular end of helix VII is in the inactive outward-facing conformation with Y^{7.53} pointing towards helix II rather than the centre of the bundle. However, in CCR5, Y^{5.58} is oriented towards the centre of the bundle, whereas in the present CCR2 structure, it faces the lipid and is sterically blocked from approaching R^{3.50} and Y^{7.53} by F^{6.38} (Fig. 3d, e). The net result of these interactions is that the crystallographically observed conformation of CCR2 appears to be even more inactive than that of CCR5 and most similar to dark rhodopsin²⁸ and Fab-bound β₂AR²⁹. Although receptor construct engineering appears to contribute to stabilization of this inactive state, the ligand binding and thermal denaturation data suggest that the concerted action of the two antagonists is also important. By directly interacting with the conserved activation microswitch residues, CCR2-RA-[R] is perfectly positioned to stabilize this inactive state: it sterically blocks Y^{7.53} from populating the active conformation and is propped against R^{3.50}, restricting its orientation away from the G-protein interface (Fig. 3b). Although located 30 Å away, BMS-681 appears to cooperate with CCR2-RA-[R] through their common interactions with helix VII, which moves outwards on the intracellular side (opposite to its movement during activation) and inwards on the extracellular side (relative to CCR5 and US28) (Fig. 3f).

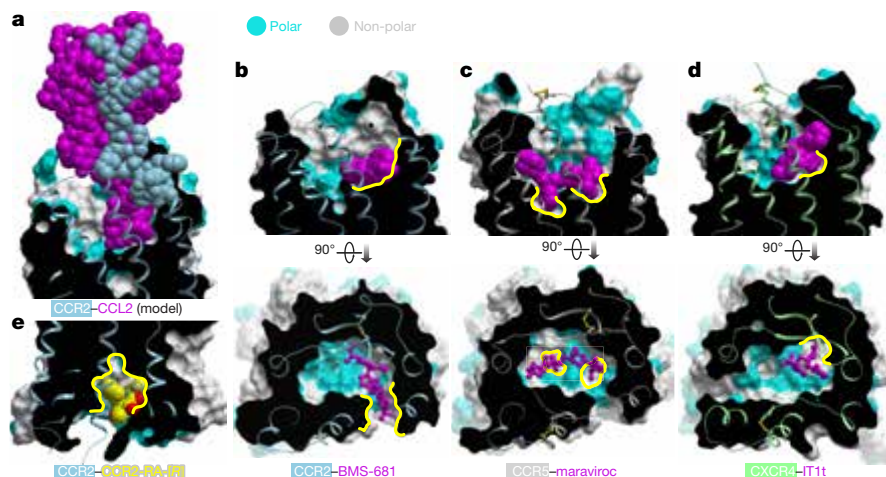


Figure 4 | Structural motifs exploited by small molecule antagonists of chemokine receptors. Receptor surface meshes are coloured by polarity (cyan, polar; grey, nonpolar). a, Modelled CCR2–CCL2 complex illustrates the extensive receptor–chemokine interface. b–d, Structures CCR2–BMS-681 (b), CCR5–maraviroc (PDB accession number 4MBS) (c), and CXCR4–IT1t (PDB accession number 3ODU) (d). Compounds utilize unique non-polar subpockets (yellow contours) within the open polar binding pockets of their target receptors. e, The allosteric pocket possesses a balanced combination of hydrophobic and polar features, making it a promising target for drug development.

The CCR2 structure has general implications for the design of drugs targeting chemokine receptors as a family. As with most protein–protein interfaces, the orthosteric binding pockets of chemokine receptors are large, wide open, and highly polar. Chemokines explore numerous hotspots within these pockets and their binding is additionally reinforced by the interaction with the flexible N termini of the receptors^{11,12} (Fig. 4a), collectively making for an extensive and versatile interaction that is conceptually difficult to inhibit with small molecules. The structure of CCR2 with BMS-681 and CCR2-RA-[R] extends the repertoire of ideas that can be used to overcome these obstacles. The binding mode of BMS-681 (Fig. 4b) contrasts with both the binding mode of maraviroc to CCR5 (Fig. 4c) where the ligand spans the major and the minor subpockets of the receptor, and that of IT1t to CXCR4 (Fig. 4d) where the ligand is entirely accommodated in the minor subpocket. While occupying the minor subpocket of CCR2, BMS-681 protrudes between helices I and VII towards the lipid bilayer (Fig. 4b) in an interaction facilitated by the trifluoromethyl group that is often present in CCR2 antagonists³⁰. This interaction enables hydrophobic anchoring of BMS-681 to the otherwise polar and open binding site of CCR2; by doing so, it parallels the role of other unique non-polar subpockets exploited by crystallized small molecule antagonists of CCR5 and CXCR4 (Fig. 4b–d). The novel subpocket explored by BMS-681 may have an additional advantage of disrupting the chemokine-compatible conformation of the receptor N terminus (Extended Data Fig. 2).

CCR2-RA-[R] demonstrates a previously unseen binding mode within an allosteric pocket on the intracellular side of CCR2. Although relatively small, this pocket has a desirable balance of polarity and hydrophobicity (Figs 2e, 4e). Homologous pockets may be present in other chemokine receptors, owing to a conserved G^{8,47}; in fact, compound binding in homologous regions has been indirectly demonstrated for CCR1 and CCR5, and directly for CCR4 (ref. 31), CXCR1, and CXCR2 (ref. 32). In most other receptors that have been crystallized thus far, the non-glycine residue at position 8.47 appears to both reduce the pocket volume and block access to the backbone amides of helix 8; consequently, the homologous pockets in these receptors may not be druggable although negative allosteric modulation with antibodies and nanobodies targeting the same region has been reported^{24,25}. By simultaneously competing with G protein and blocking activation-related conformational changes, compound binding in the allosteric pocket seems a powerful way to antagonize the receptor. Therefore, for receptors in which the allosteric pocket is druggable, targeting it with small molecules may open new avenues for GPCR drug discovery.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 June; accepted 7 November 2016.

Published online 7 December 2016.

- Scholten, D. J. *et al.* Pharmacological modulation of chemokine receptor function. *Br. J. Pharmacol.* **165**, 1617–1643 (2012).
- O'Connor, T., Borsig, L. & Heikenwalder, M. CCL2-CCR2 signaling in disease pathogenesis. *Endocr. Metab. Immune Disord. Drug Targets* **15**, 105–118 (2015).
- Lim, S. Y., Yuzhalin, A. E., Gordon-Weeks, A. N. & Muschel, R. J. Targeting the CCL2-CCR2 signaling axis in cancer metastasis. *Oncotarget* **7**, 28697–28710 (2016).
- Solari, R., Pease, J. E. & Begg, M. "Chemokine receptors as therapeutic targets: why aren't there more drugs?". *Eur. J. Pharmacol.* **746**, 363–367 (2015).
- Cooke, R. M., Brown, A. J., Marshall, F. H. & Mason, J. S. Structures of G protein-coupled receptors reveal new opportunities for drug discovery. *Drug Discov. Today* **20**, 1355–1364 (2015).
- Carter, P. H. *et al.* Discovery of a potent and orally bioavailable dual antagonist of CC chemokine receptors 2 and 5. *ACS Med. Chem. Lett.* **6**, 439–444 (2015).
- Dasse, O. *et al.* Novel, acidic CCR2 receptor antagonists: lead optimization. *Lett. Drug Des. Discov.* **4**, 263–271 (2007).
- Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4**, 706–731 (2009).
- Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330**, 1066–1071 (2010).
- Tan, Q. *et al.* Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* **341**, 1387–1390 (2013).

- Qin, L. *et al.* Structural biology. Crystal structure of the chemokine receptor CXCR4 in complex with a viral chemokine. *Science* **347**, 1117–1122 (2015).
- Burg, J. S. *et al.* Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor. *Science* **347**, 1113–1117 (2015).
- Montecarlo, F. S. & Charo, I. F. The amino-terminal domain of CCR2 is both necessary and sufficient for high affinity binding of monocyte chemoattractant protein 1. Receptor activation by a pseudo-tethered ligand. *J. Biol. Chem.* **272**, 23186–23190 (1997).
- Zhang, K. *et al.* Structure of the human P2Y12 receptor in complex with an antithrombotic drug. *Nature* **509**, 115–118 (2014).
- Weik, M. *et al.* Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proc. Natl Acad. Sci. USA* **97**, 623–628 (2000).
- Cherney, R. J. *et al.* Discovery of disubstituted cyclohexanes as a new class of CC chemokine receptor 2 antagonists. *J. Med. Chem.* **51**, 721–724 (2008).
- Berkhout, T. A. *et al.* CCR2: characterization of the antagonist binding site from a combined receptor modeling/mutagenesis approach. *J. Med. Chem.* **46**, 4070–4086 (2003).
- Hall, S. E. *et al.* Elucidation of binding sites of dual antagonists in the human chemokine receptors CCR2 and CCR5. *Mol. Pharmacol.* **75**, 1325–1336 (2009).
- Cherney, R. J. *et al.* Synthesis and evaluation of cis-3,4-disubstituted piperidines as potent CC chemokine receptor 2 (CCR2) antagonists. *Bioorg. Med. Chem. Lett.* **18**, 5063–5065 (2008).
- Zweemer, A. J. *et al.* Discovery and mapping of an intracellular antagonist binding site at the chemokine receptor CCR2. *Mol. Pharmacol.* **86**, 358–368 (2014).
- Blankenship, E., Vahedi-Faridi, A. & Lodowski, D. T. The high-resolution structure of activated opsin reveals a conserved solvent network in the transmembrane region essential for activation. *Structure* **23**, 2358–2364 (2015).
- Rasmussen, S. G. *et al.* Crystal structure of the β_2 adrenergic receptor–Gs protein complex. *Nature* **477**, 549–555 (2011).
- Zweemer, A. J. *et al.* Multiple binding sites for small-molecule antagonists at the CC chemokine receptor 2. *Mol. Pharmacol.* **84**, 551–561 (2013).
- Hino, T. *et al.* G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. *Nature* **482**, 237–240 (2012).
- Staus, D. P. *et al.* Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. *Nature* **535**, 448–452 (2016).
- Katritch, V., Cherezov, V. & Stevens, R. C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* **53**, 531–556 (2013).
- Manglik, A. *et al.* Structural insights into the dynamic process of β_2 -adrenergic receptor signaling. *Cell* **161**, 1101–1111 (2015).
- Palczewski, K. *et al.* Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**, 739–745 (2000).
- Rasmussen, S. G. *et al.* Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383–387 (2007).
- Pease, J. & Horuk, R. Chemokine receptor antagonists. *J. Med. Chem.* **55**, 9363–9392 (2012).
- Andrews, G., Jones, C. & Wreggett, K. A. An intracellular allosteric site for a specific class of antagonists of the CC chemokine G protein-coupled receptors CCR4 and CCR5. *Mol. Pharmacol.* **73**, 855–867 (2008).
- Nicholls, D. J. *et al.* Identification of a putative intracellular allosteric antagonist binding-site in the CXC chemokine receptors 1 and 2. *Mol. Pharmacol.* **74**, 1193–1202 (2008).

Acknowledgements We thank A. Ishchenko and H. Zhang for help with X-ray data collection, C. Wang and H. X. Wu for suggestions on construct design, F. Li for help with data processing, and M. Galella for assistance with BMS compound data and statistics. We thank C. Ogata, R. Sanishvili, N. Venugopalan, M. Becker, and S. Corcoran at beamline 23ID at GM/CA CAT Advanced Photon Source. Funding for this research was provided by National Institutes of Health grants R01 GM071872, U54 GM094618, R01 AI118985, R21 AI121918, and R21 AI122211. GM/CA@APS has been funded in whole or in part with federal funds from the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006). This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract number DE-AC02-06CH11357.

Author Contributions I.K. and T.M.H. designed the study and coordinated all experiments. Y.Z. designed and engineered protein constructs, performed crystallization experiments, collected the diffraction data, and determined the structure. L.Q., M.G., and C.Z. assisted with protein engineering and crystallization. G.W.H. assisted with structure determination and refinement. A.P.I. and L.H.H. designed, and N.V.O.Z. and H.d.V. performed, equilibrium and kinetics binding experiments. I.K. performed computational and bioinformatics analyses. R.J.C., P.C., and A.T. synthesized, characterized, and crystallized the BMS compound analogues. M.D. assisted with compound crystallization. D.S. assisted with the allosteric compound characterization. R.A. assisted with structure analysis. V.C. and R.C.S. assisted with crystallization. Y.Z., N.V.O.Z., A.P.I., L.H.H., I.K., and T.M.H. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.M.H. (thandel@ucsd.edu), I.K. (ikufareva@ucsd.edu) or L.H.H. (l.h.heitman@lacr.leidenuniv.nl).

METHODS

Design and expression of CCR2-T4L fusion constructs. The sequence of human CCR2 isoform B (Uniprot ID P41597-2) was engineered for crystallization by truncation of C-terminal residues 329–360 and by grafting T4 lysozyme (T4L) into the ICL3. In the process of construct optimization, the native CCR2 residues between L226^{5,62} and R240^{6,32} (L226^{5,62}-KTLRCRNEKKRH-R240^{6,32}) were removed and replaced with corresponding residues from the crystallized structure of M2 muscarinic acetylcholine receptor (PDB accession number 3UON, resulting amino-acid sequence S226^{5,62}-RASKSRI-T4L-PPPSREK-K240^{6,32}). The presence of T4L in ICL3 is expected to prevent receptor activation; however the similar affinities of BMS-681 and CCR2-RA-[R] for both wild-type (WT) CCR2 and CCR2-T4L (Extended Data Fig. 4a, b and Extended Data Table 3) suggest that the fusion construct is a good surrogate of WT CCR2 for understanding ligand recognition.

The CCR2-T4L coding sequence was cloned into a modified pFastBac1 vector (Invitrogen) with an HA signal sequence followed by a Flag tag at the N terminus and a PreScission protease site followed by a 10× His tag and another Flag tag at the C terminus. The receptor was expressed in *Spodoptera frugiperda* (Sf9) cells. High-titre recombinant baculovirus (>10⁹ viral particles per millilitre) was obtained using the Bac-to-Bac Baculovirus Expression System (Invitrogen) as previously described¹¹. Sf9 cells at a cell density of (2–3) × 10⁶ cells per millilitre were infected with P1 virus at a multiplicity of infection of 5. Cells were harvested by centrifugation 48 h after infection and stored at –80 °C until use.

Purification of CCR2-T4L. Insect cell membranes were prepared by thawing frozen cell pellets in a hypotonic buffer containing 10 mM HEPES (pH 7.5), 10 mM MgCl₂, 20 mM KCl, and EDTA-free complete protease inhibitor cocktail tablets (Roche). Extensive washing of the raw membranes was performed by repeated douncing and centrifugation in the same hypotonic buffer (two or three times) and then in a high osmotic buffer containing 1.0 M NaCl, 10 mM HEPES (pH 7.5), 10 mM MgCl₂, 20 mM KCl, and EDTA-free complete protease inhibitor cocktail tablets (three or four times), thereby separating soluble and membrane associated proteins from integral transmembrane proteins. Stock solutions (40 mM) of BMS-681 and CCR2-RA-[R] were made in isopropanol. Washed membranes were resuspended into a buffer containing 50 μM BMS-681, 2 mg/ml iodoacetamide, and EDTA-free complete protease inhibitor cocktail tablets, and incubated at 4 °C for 1 h before solubilization. The membranes were then solubilized in 50 mM HEPES (pH 7.5), 400 mM NaCl, 1% (w/v) *n*-dodecyl-β-D-maltopyranoside (DDM, Anatrace), 0.2% (w/v) cholesteryl hemisuccinate (CHS, Sigma) at 4 °C for 3 h. The supernatant was isolated by centrifugation at 50,000g for 30 min, and incubated in 20 mM HEPES (pH 7.5), 400 mM NaCl with TALON IMAC resin (Clontech) overnight at 4 °C. After binding, the resin was washed without addition of ligands with ten column volumes of Wash I Buffer (50 mM HEPES (pH 7.5), 400 mM NaCl, 10% (v/v) glycerol, 0.1% (w/v) DDM, 0.02% (w/v) CHS, 10 mM imidazole), followed by four column volumes of Wash II Buffer (50 mM HEPES (pH 7.5), 400 mM NaCl, 10% (v/v) glycerol, 0.02% (w/v) DDM, 0.01% (w/v) CHS, 50 mM imidazole). The protein was then eluted with three to four column volumes of Elution Buffer (50 mM HEPES (pH 7.5), 1 μM BMS-681, 400 mM NaCl, 10% (v/v) glycerol, 0.02% (w/v) DDM, 0.01% (w/v) CHS, 250 mM imidazole). PD MiniTrap G-25 columns (GE Healthcare) were used to remove imidazole. The protein was then treated overnight with His-tagged PreScission protease to cleave the C-terminal His-tag and Flag-tag. PreScission protease and the cleaved C-terminal fragment were removed by binding to TALON IMAC resin for 2 h at 4 °C. The protein was collected as the TALON IMAC column flow-through. The protein was supplemented with 75 μM each of BMS-681 and CCR2-RA-[R] before being concentrated to 30 mg/ml with a 100 kDa molecular mass cut-off Amicon centrifuge concentrator (Millipore). The estimated final compound concentrations were ~1–2 mM for both compounds.

Protein stability assays. The thermostability of CCR2-T4L was analysed by a differential scanning fluorimetry assay adapted from previous publications³³ using a RotorGene Q 6-plex RT-PCR machine (Qiagen). Briefly, 1–5 μg of protein was mixed with 3 μM 7-diethylamino-3-(4'-maleimidylphenyl)-4-methylcoumarin (CPM) dye (2.5 mM stock in DMSO) in 25 mM HEPES pH 7.5, 400 mM NaCl, 0.02% DDM, 0.004% CHS, 10% glycerol, and indicated concentrations of compounds to a final volume of 20 μl; samples were incubated for 5 min at room temperature and then heated gradually from 28 °C to 90 °C at a rate of 0.8 °C/min, with CPM fluorescence (excitation 365 nm, emission 460 nm) recorded every 1 °C. The melting temperature (*T*_m) was determined from the first derivative of the denaturation curve, using Rotor-Gene Q – Pure Detection software (version 2.0.3).

Crystallization. Purified CCR2 in complex with BMS-681 and CCR2-RA-[R] was reconstituted into LCP by mixing with molten lipid using a mechanical syringe mixer⁸. The protein–LCP mixture contained 40% (w/w) receptor solution, 54% (w/w) monoolein, and 6% (w/w) cholesterol. Crystallization trials were performed

in 96-well glass sandwich plates (Hampton research) using a Mosquito LCP robot (TTP Labtech) by dispensing 45 nl of protein-laden LCP and 800 nL of precipitant solution per well. Plates were incubated and imaged at 20 °C. Initial crystal hits were found from a precipitant condition containing 100 mM MES, pH 6.5, 30% (v/v) PEG400, 100 mM Li₂SO₄. After optimization, diffraction-quality crystals were obtained from 100 mM MES, pH 6.5, 30–32% (v/v) PEG400, 75–85 mM Li₂SO₄. Crystals usually grew to a maximum size of 60 μm × 10 μm × 10 μm in 1 week, and were harvested directly from the LCP matrix using MiTeGen micromounts and flash cooled in liquid nitrogen.

Data collection and structure determination. X-ray diffraction data were collected using a 10 μm collimated minibeam at a wavelength of 1.0332 Å with a Pilatus3 6M direct detector on the 23ID-D beamline (GM/CA CAT) of the Advanced Photon Source at the Argonne National Laboratory. Crystals were located and aligned by the rastering strategy³⁴. Among the several hundred crystal samples screened, most crystals diffracted to 2.8–3.5 Å resolution when exposed to 0.3 s of unattenuated beam using 0.3° oscillations. A 93.1% complete data set at 2.80 Å resolution was obtained by merging data from 17 crystals, using XDS³⁵ and Aimless³⁶. As the data showed anisotropy, the UCLA Diffraction Anisotropy Server (<http://services.mbi.ucla.edu/anisotropy/>) was used to truncate the data to 3.0 Å along both *a** and *b** axes, and to 2.81 Å along the *c** axis. Initial phase information was obtained by molecular replacement with the program Phaser³⁷ using the receptor portion of the CCR5 structure (PDB accession number 4MBS) converted to polyalanines, and the T4L portion of the CXCR4 structure (PDB accession number 3ODU) as search models. The correct molecular replacement solution (translation function Z-score = 14.8) contained one CCR2-T4L molecule in the asymmetric unit. Refinement was performed with Phenix³⁸ followed by manual examination and rebuilding of the refined coordinates in the program COOT³⁹ using both $|2F_o| - |F_c|$ and $|F_o| - |F_c|$ maps, as well as omit maps. The final model included 295 residues (37–225 and 241–319) of the 360 residues of CCR2 and residues 2–161 of T4L plus 16 residues of two 8-residue linkers. The remaining N- and C-terminal residues were disordered and were not built. Strong electron density for one metal ion was observed. The identity of the ion was determined to be Zn²⁺ by X-ray fluorescence scans (Extended Data Fig. 5). The zinc ion is coordinated by a water molecule as well as side chains of H144^{3,56}, E238, and E1005. Data collection and refinement statistics are shown in Extended Data Table 1.

Crystallization and structure determination of BMS-681. BMS-681 was dissolved in a minimal amount of CH₃CN and then 15% water was added. After standing overnight, the resulting crystals were collected. Data were obtained on a Bruker-AXS X8-Proteum Kappa goniometer and APEXII detector. Intensities were measured using Cu Kα radiation ($\lambda = 1.5418$ Å) with the crystal kept at a constant temperature using an Oxford cryo system during data collection. Indexing and processing of the measured intensity data were performed with the SAINT-APEX2 (Bruker-AXS) program suite, structure solution with SHELXS-97, and structure refinement with SHELXL-97.

The derived atomic parameters (coordinates and temperature factors) were refined through full matrix least-squares. The function minimized in the refinements was $\sum_w (|F_o| - |F_c|)^2$. *R* is defined as $\sum ||F_o| - |F_c|| / \sum |F_o|$ while $R_w = [\sum_w (|F_o| - |F_c|)^2 / \sum_w |F_o|^2]^{1/2}$ where *w* is an appropriate weighting function based on errors in the observed intensities. Hydrogens were introduced in idealized positions with isotropic temperature factors, but no hydrogen parameters were varied. It should be noted that the refinement model illustrates disorder and partial occupancy factors of 'guest' solvent/water molecules within the crystalline lattice. The atomic positions of these disordered molecules were taken from the difference map analysis, which showed peaks of electron density of varying intensities at the refined positions representing the disordered solvent/water molecules. Data collection and refinement statistics are shown in Extended Data Table 2.

Cell culture and transfections. Chinese hamster ovary (CHO) cells (provided by H. den Dulk, Leiden University, The Netherlands; originally obtained from and certified by American Type Culture Collection) were cultured in Dulbecco's Modified Eagle Medium/F-12 Nutrient Mixture (DMEM/F-12) supplemented with 10% (v/v) newborn calf serum, 50 IU/ml penicillin, and 50 μg/ml streptomycin; they were maintained at 37 °C and in 5% CO₂. Cells were subcultured twice a week at a ratio of 1:30 to 1:50 by trypsinization. Transient transfection of CHO cells with WT CCR2 and CCR2-T4L constructs was performed using a polyethylenimine method, as described previously²³. Briefly, CHO cells were grown on plates (diameter 15 cm) to around 50% confluence and then transfected with a DNA/polyethylenimine mixture containing 10 μg plasmid DNA—previously diluted in 150 mM NaCl solution—mixed with polyethylenimine solution (1 mg/ml) at a 1:6 DNA:polyethylenimine mass ratio. Before adding 1 ml of the transfection mixture to each plate, the culture medium of the cells was refreshed and the mixture incubated for 20 min at room temperature. Following transfection, cells were incubated for 48 h at 37 °C and 5% CO₂ before membrane preparation.

Twenty-four hours after transfection, sodium butyrate was added to each plate at a final concentration of 3 mM to increase receptor expression. CHO cells were tested for mycoplasma contamination before use, the outcome of which was negative.

Membrane preparation. Membranes from CHO cells transiently expressing the WT CCR2 or CCR2-T4L were prepared as described previously²³. Briefly, cells were detached from plates (diameter 15 cm) using 5 ml of phosphate-buffered saline and centrifuged for 5 min at 3000g. The membranes were separated from the cytosolic fractions by several centrifugation and homogenization steps. First, the pellets were resuspended and homogenized in ice-cold membrane buffer (50 mM Tris-HCl buffer, supplemented with 5 mM MgCl₂, pH 7.4) using an Ultra Thurrax Homogenizer (IKA-Werke, Staufen, Germany). Homogenized membranes were then centrifuged in an Optima LE-80 K ultracentrifuge (Beckman Coulter, Fullerton, California, USA) at 31,000g for 20 min at 4 °C. The final membrane pellet was resuspended also in ice-cold membrane buffer and aliquoted before storage. Membrane aliquots were stored at −80 °C and protein concentrations were measured using a standard BCA protein determination assay (Pierce Chemical Company, Rockford, Illinois, USA).

Radioligand binding assays. [³H]INCB-3344 (specific activity 32 Ci mmol^{−1}) and [³H]CCR2-RA (specific activity 63 Ci mmol^{−1}) were custom-labelled by Vitrac (Placentia, California, USA). JNJ-27141491 was synthesized as described previously⁴⁰. INCB-3344 and CCR2-RA-[R] were synthesized in-house as described previously^{7,41}.

All radioligand binding assays were performed at 25 °C in a 100 µL reaction volume containing assay buffer (50 mM Tris-HCl buffer (pH 7.4), 5 mM MgCl₂, 0.1% CHAPS) and 30 µg of membrane protein from CHO cells transiently expressing WT CCR2 or CCR2-T4L. For competition binding assays with [³H]INCB-3344, a concentration of 5 nM [³H]INCB-3344 was used, and non-specific binding was determined with 10 µM of unlabelled INCB-3344. In the case of [³H]CCR2-RA competition binding assays, a radioligand concentration of 3 nM was used and non-specific binding was determined with 10 µM of JNJ-27141491. In all cases, homologous or competition displacement assays were performed using six increasing concentrations of competing ligands. Kinetic experiments were also performed at 25 °C using 7 nM [³H]CCR2-RA and 30 µg of membrane protein in a 100 µL reaction volume. For association experiments, CHO-CCR2 or CHO-CCR2-T4L membranes were added to the reaction at eight different time points, in the absence or presence of 1 µM BMS-681. For dissociation experiments, membranes were first incubated with radioligand for 90 min; dissociation was then initiated by addition of 10 µM of CCR2-RA-[R] at 12 different time points, in the presence or absence of 1 µM BMS-681. More time points were used in the dissociation assays, to characterize the biphasic profile of [³H]CCR2-RA dissociation. In all cases, total radioligand binding did not exceed 10% of the total radioligand added to avoid ligand depletion. For all experiments, incubation was terminated by dilution with ice-cold wash buffer (50 mM Tris-HCl buffer (pH 7.4), 5 mM MgCl₂, 0.05% CHAPS). Separation of bound from free radioligand was achieved by rapid filtration over a 96-well GF/B filter plate using a Perkin Elmer Filtermate-harvester (Perkin Elmer, Groningen, The Netherlands) and filter-bound radioactivity was

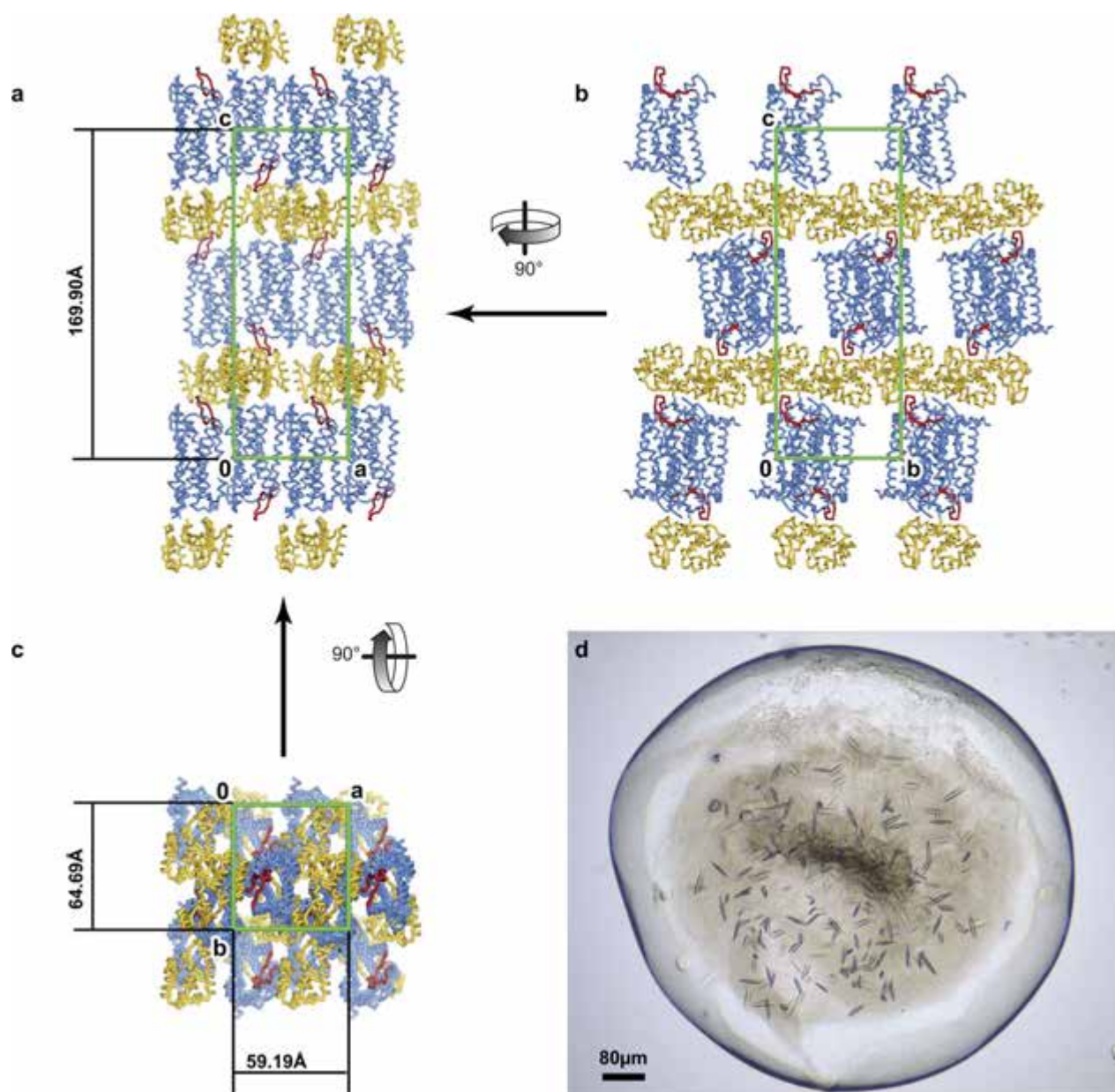
determined in a Perkin Elmer 2450 Microbeta2 plate counter after addition of 25 µL Microscint scintillation cocktail per well (Perkin-Elmer, Groningen, The Netherlands).

Statistical methods. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

All radioligand binding data were analysed using Prism 6.0 and 7.0 (GraphPad Software, San Diego, California, USA). The pIC₅₀ values were obtained by nonlinear regression analysis of competition displacement assays. Apparent association rate constants (*k*_{obs}) and maximum binding (*B*_{max}, used to calculate %*B*/*B*_{control}) were determined by fitting the association data to a one-phase exponential association function. Dissociation rate constants were determined by fitting the dissociation data to a monophasic (*k*_{off}) or biphasic (*k*_{off, fast} and *k*_{off, slow}) exponential decay model. All data shown represent means ± s.e.m. of at least three independent experiments performed in duplicate. An unpaired, two-tailed Student's *t*-test was used to compare differences in pIC₅₀ as well as differences in kinetic parameters. Differences in binding enhancement (%Binding) in the absence (set at 100%) or presence of BMS-681 were analysed using a one-way analysis of variance with Dunnett's post-hoc test. Significant differences are denoted as follows: **P* < 0.05, ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001.

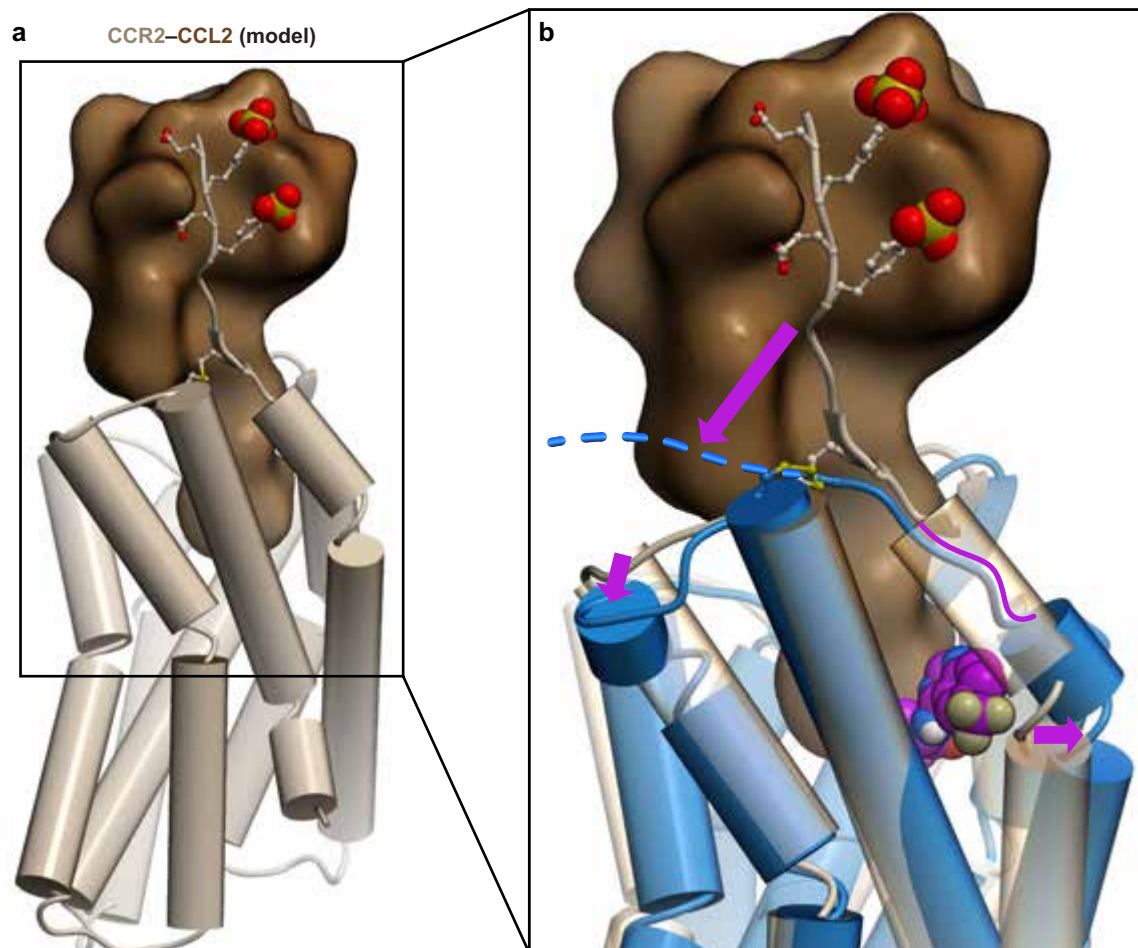
Data availability. The atomic coordinates and structure factors for the CCR2–BMS-681–CCR2-RA-[R] complex have been deposited in the Protein Data Bank under accession number 5T1A. The structure of free BMS-681 is deposited in the Cambridge Crystallographic Data Centre (<http://www.ccdc.cam.ac.uk/>) under accession number 1479580. All other data are available from the corresponding authors upon reasonable request.

33. Alexandrov, A. I., Mileni, M., Chien, E. Y. T., Hanson, M. A. & Stevens, R. C. Microscale fluorescent thermal stability assay for membrane proteins. *Structure* **16**, 351–359 (2008).
34. Cherezov, V. *et al.* Rastering strategy for screening and centring of microcrystal samples of human membrane proteins with a sub-10 µm size X-ray synchrotron beam. *J. R. Soc. Interface* **6** (Suppl. 5), S587–S597 (2009).
35. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
36. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
37. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
38. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
39. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
40. Doyon, J. *et al.* Discovery of potent, orally bioavailable small-molecule inhibitors of the human CCR2 receptor. *ChemMedChem* **3**, 660–669 (2008).
41. Brodmerkel, C. M. *et al.* Discovery and pharmacological characterization of a novel rodent-active CCR2 antagonist, INCB3344. *J. Immunol.* **175**, 5370–5378 (2005).
42. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).



Extended Data Figure 1 | CCR2-T4L crystals and crystal packing. **a–c**, Crystal packing of CCR2-T4L. CCR2 is a blue ribbon with ECL2 coloured red and T4L yellow. The unit cell is shown as a green box. CCR2-T4L molecules are arranged in a type I packing with hydrophilic stacking mediated by T4L and T4L-ECL2 interactions along axis *c*. **a**, Crystal packing in the *ac* plane. CCR2 makes abundant hydrophobic contacts with its neighbour via an interface mediated by antiparallel helix IV-helix VI interactions related by a screw axis along axis *a*. **b**, Crystal

packing in the *bc* plane. Contacts between receptors and T4L involve ECL2 and the intracellular surface of CCR2 including helix VIII. Direct contacts between T4L are along axis *b*. One layer of CCR2-T4L molecules at the very top of the stacking column is omitted for clarity. **c**, Crystal packing in the *ab* plane. There are no direct interactions between T4L along axis *a*. **d**, Crystals of CCR2-T4L in the LCP bolus. Average crystals grew to 60 μm × 10 μm × 10 μm before harvesting.

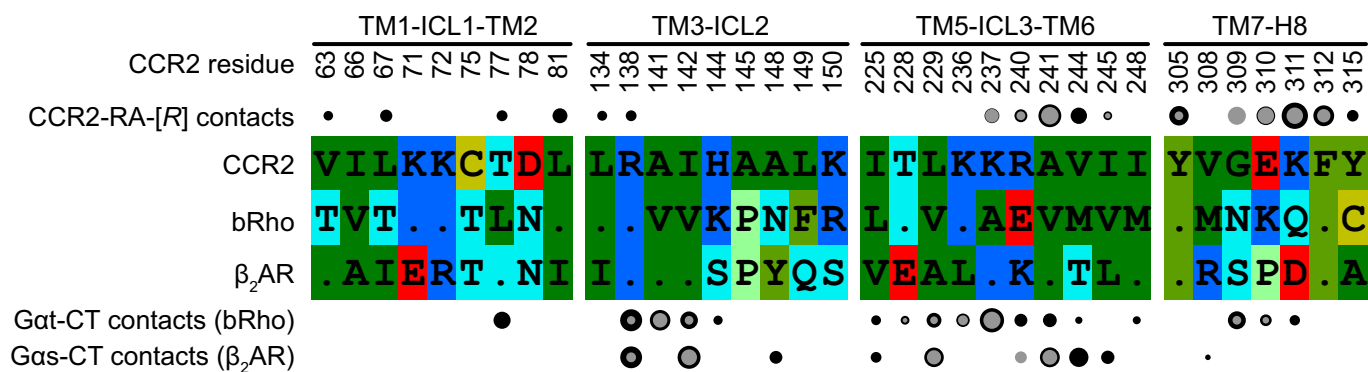


CCR2-CCL2 (model) vs CCR2-BMS-681 (structure extended with NT)

Extended Data Figure 2 | BMS-681 binding may disrupt a chemokine-recognizing conformation of the CCR2 N terminus and helix I.

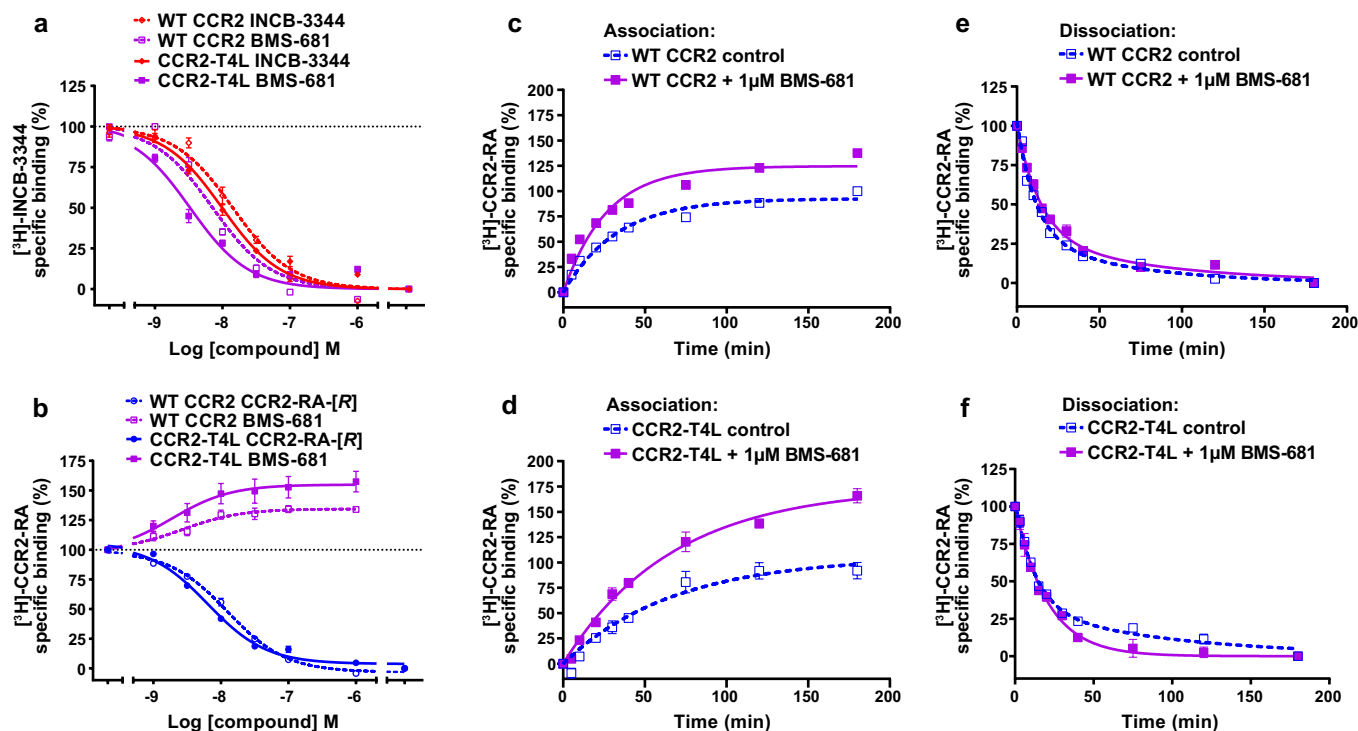
a, Model of CCR2-CCL2 built by homology from the structure of CXCR4-vMIP-II¹¹ suggests that a productive chemokine-compatible conformation of the receptor requires re-orientation of the N terminus

from almost parallel to almost perpendicular to the membrane plane, and formation of an extra helical turn in helix I to bring it closer to helix VII and ECL3. **b**, Binding of BMS-681 may disrupt this chemokine-compatible conformation by inserting between helices I and VII.



Extended Data Figure 3 | CCR2-RA-[R] directly binds to CCR2 residues that are homologous to those involved in G-protein coupling in other GPCRs. Partial alignment of intracellular regions of CCR2 and homologous regions in bovine Rho (bRho) and β_2 adrenergic receptor (β_2 AR), alongside profile of contacts that CCR2-RA-[R], the $G\alpha_t$ C-terminal peptide²¹, and $G\alpha_s$ C terminus²² make with the three

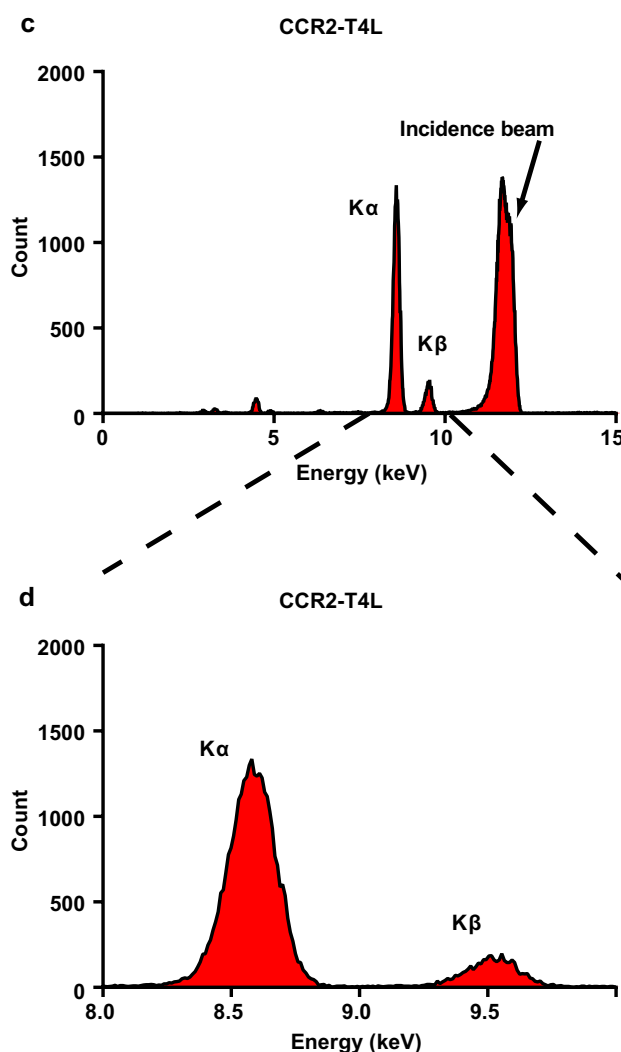
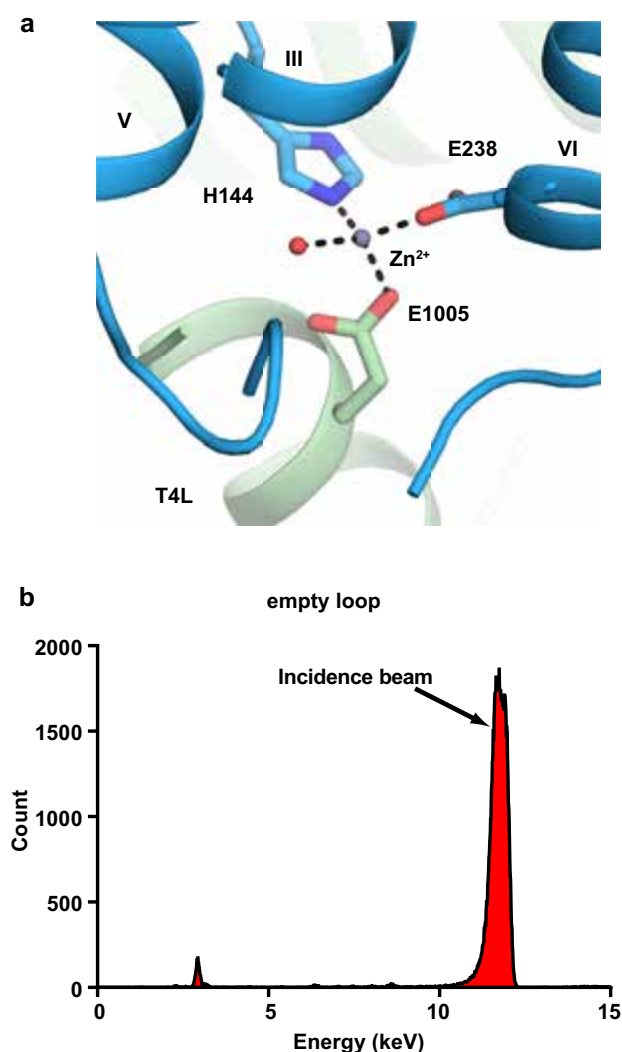
respective receptors. Contacts are shown by circles above and below the alignment, with circle area indicative of contact strength. Backbone and side-chain contacts are grey and black, respectively. Assuming structural homology between the CCR2-G-protein interface and at least one of the bRho- $G\alpha_t$ and β_2 AR- $G\alpha_s$ interfaces, several residue positions seem to be involved in binding both CCR2-RA-[R] and the C terminus of the G protein.



Extended Data Figure 4 | Equilibrium binding and binding kinetics of BMS-681 and CCR2-RA-[R] with WT CCR2 and CCR2-T4L.

a, b, Displacement of $[^3\text{H}]$ INCB-3344 (5 nM, **a**) and $[^3\text{H}]$ CCR2-RA-[R] (3 nM, **b**) from WT CCR2 and CCR2-T4L in CHO cells by increasing concentrations of unlabelled INCB-3344, CCR2-RA-[R] and BMS-681. **c, d**, Association and (**e, f**) dissociation of 7 nM $[^3\text{H}]$ CCR2-RA from CHO

cell membranes transiently expressing WT CCR2 (**c, e**) or CCR2-T4L (**d, f**) at 25°C, in the absence or presence of 1 μM BMS-681. Figures represent normalized and combined data from three independent experiments performed in duplicate, with results presented as mean \pm s.e.m. percentage of specific $[^3\text{H}]$ CCR2-RA binding.



Extended Data Figure 5 | A Zn^{2+} binding site was identified by X-ray fluorescence emission analysis of the CCR2-T4L-BMS-681-CCR2-RA-[R] crystals. a, View of the Zn^{2+} ion at an interface formed by CCR2 helices III and VI and the N terminus of T4L. The Zn^{2+} ion is coordinated by side chains of H144^{3,56} (from WT receptor), E238^{6,30} (from the engineered part of the receptor), and E1005 (from T4L) as well as a structured water. **b**, Background fluorescence signal of an

empty MiTeGen micromount is low, indicating the absence of metal ion. Excitation at 12 keV results in a peak at 11.7 keV (owing to the incidence beam). **c**, X-ray fluorescence emission signal from a wide fluorescence scan of the CCR2-T4L crystal. The fluorescence peaks at 8.60 keV and 9.53 keV correspond to X-ray emission lines $K\alpha$ (8.64 keV) and $K\beta$ (9.57 keV) and indicate the presence of Zn^{2+} bound to CCR2-T4L. **d**, A zoomed-in view of the X-ray fluorescence emission signal from **c**.

Extended Data Table 1 | Data collection and refinement statistics (molecular replacement)

CCR2-T4L-BMS-681-CCR2-RA-[R] [†]	
Data collection [‡] wavelength (Å)	1.03319
Space group	P2 ₁ 2 ₁ 2 ₁
Unit cell parameters <i>a,b,c</i> (Å)	59.19 64.69 169.90
Number of reflections measured	82,111
Number of unique reflections	15,550
Resolution (Å)	48-2.8 (2.95-2.8)
<i>R</i> _{merge} (%)	22.5(101)
<i>R</i> _{pim} (%)	12.8(88.4)
Mean <i>I</i> / <i>s(I)</i>	6.9(0.8)
Completeness (%)	93.1(66.6)
Redundancy	5.3(1.8)
Refinement	
Resolution (Å)	25-2.81 (3.0, 3.0, 2.81)
Number of reflections (test set)	14515 (746)
<i>R</i> _{work} / <i>R</i> _{free}	0.233/0.274 (0.319/0.392)
Number of atoms	
CCR2	3,580
T4L	2,215
BMS-681	1,243
CCR2-RA-[R]	35
Monoolein	24
Sulfate	25
Water	20
Zn	17
Mean overall B value (Å ²)	
Wilson B	41.4
Protein	40.4
Ligands	41.5
Water	41.3
Root mean square deviation	
Bond lengths (Å)	22.9
Bond angles (°)	0.003
Ramachandran plot statistics [§] (%)	
Favored regions	0.85
Allowed regions	97.1
Disallowed regions	2.9
	0

[†]Diffraction data from 17 crystals were merged into a complete data set.

[‡]Highest resolution shell statistics are shown in parentheses.

[§]As defined in MolProbity⁴².

Extended Data Table 2 | Small-molecule (BMS-681) X-ray data collection and refinement

Empirical formula	C₂₆ H₃₆ F₃ N₅ O_{3.58}
Formula weight	532.80
Temperature	173(2) K
Wavelength	1.54178 Å
Crystal system	Tetragonal
Space group	P ₄ ₃ 2 ₁ 2
Unit cell dimensions	a = 20.4436(4) Å α = 90°. b = 20.4436(4) Å β = 90°. c = 28.9325(7) Å γ = 90°.
Volume	12092.1(4) Å ³
Z	16
Density (calculated)	1.171 Mg/m ³
Absorption coefficient	0.768 mm ⁻¹
F(000)	4522
Crystal size	0.46 x 0.18 x 0.16 mm ³
Theta range for data collection	2.65 to 58.78°.
Resolution range	16.7 to 0.9 Å
Index ranges	-22 ≤ h ≤ 22, -21 ≤ k ≤ 22, -31 ≤ l ≤ 14
Reflections collected	108743
Independent reflections	8518 [R(int) = 0.1259]
Completeness to theta = 58.78°	98.6 %
Absorption correction	None
Refinement method	Full-matrix least-squares on F ²
Data / restraints / parameters	8518 / 22 / 713
Goodness-of-fit on F²	1.058
Final R indices [I > 2σ(I)]	R1 = 0.0770, wR2 = 0.2087
R indices (all data)	R1 = 0.0860, wR2 = 0.2178
Absolute structure parameter; Flack(x)	0.1(2)
Absolute structure parameter; Hooft(y), P3true	0.03(5), 1.000
Largest diff. peak and hole	0.543 and -0.405 e.Å ⁻³

Extended Data Table 3 | Displacement of specific [³H]INCB-3344 (5 nM) and [³H]CCR2-RA (3 nM) binding from CCR2 constructs transiently expressed on CHO cells

Construct	³ H]-INCB-3344 displacement by INCB-3344	³ H]-INCB-3344 displacement by BMS-681	³ H]-CCR2-RA displacement by CCR2-RA-[R]	³ H]-CCR2-RA enhancement by BMS-681
	pIC ₅₀ ± S.E.M (IC ₅₀ , nM)			%Binding
WT CCR2	7.8 ± 0.0 (17)	8.1 ± 0.0 (8)	7.9 ± 0.0 (13)	134 ± 3% ^{†**}
CCR2-T4L	8.1 ± 0.1* (8)	8.6 ± 0.1** (3)	8.2 ± 0.0** (6)	157 ± 13% ^{†****}

Values represent mean ± s.e.m. of at least three independent experiments performed in duplicate.

[†]Percentage of [³H]CCR2-RA (3 nM) binding in presence of BMS-681 (1 μM). Values higher than 100% represent binding enhancement compared with the 100% control without BMS-681.

Differences in pIC₅₀ values between constructs were analysed using a Student's *t*-test, with significant differences noted as follows: **P* < 0.05, ***P* < 0.01.

Differences in percentage Binding in the absence (100%) and presence of BMS-681 were analysed using a one-way analysis of variance with Dunnett's post-hoc test, with significant differences noted as follows: ***P* < 0.01, *****P* < 0.0001.

Extended Data Table 4 | Observed association and dissociation rate constants of [³H]CCR2-RA (7 nM) on membranes from CHO cells transiently expressing WT CCR2 and CCR2-T4L, in the absence or presence of 1 μM BMS-681

	CHO-CCR2		CHO-CCR2-T4L	
	Control	+ 1 μM BMS-681	Control	+ 1 μM BMS-681
k_{obs} (min⁻¹)	0.031 ± 0.002	0.038 ± 0.003*	0.015 ± 0.003	0.015 ± 0.001
% B/B_{control} [†]	100 ± 0.0	135 ± 2.0****	100 ± 0.0	162 ± 8.4**
k_{off,fast} (min⁻¹)	0.089 ± 0.015	0.069 ± 0.012*	0.077 ± 0.013	0.049 ± 0.003‡
k_{off,slow} (min⁻¹)	0.016 ± 0.005	0.012 ± 0.004	0.010 ± 0.003	
%fast	70 ± 10	71 ± 11	69 ± 8	N/A‡

Values represent mean ± s.e.m. of three independent experiments performed in duplicate.

[†]The percentage of maximum binding in the absence (B_{control}) or presence (B) of BMS-681 (1 μM).

[‡]For CHO-CCR2-T4L only, dissociation kinetics of [³H]CCR2-RA (7 nM) in the presence of BMS-681 (1 μM) fitted best with a monophasic exponential decay model, resulting in a single k_{off} value, as shown in the table. Thus, for CHO-CCR2-T4L, the statistical significance between k_{off} measurements with and without BMS-681 could not be calculated.

Statistical significance was analysed using a Student's t-test, with significant differences versus control noted as follows: *P < 0.05, **P < 0.01, ****P < 0.0001.

Intracellular allosteric antagonism of the CCR9 receptor

Christine Oswald^{1*}, Mathieu Rappas^{1*}, James Kean^{1*}, Andrew S. Doré¹, James C. Errey¹, Kirstie Bennett¹, Francesca Deflorian¹, John A. Christopher¹, Ali Jazayeri¹, Jonathan S. Mason¹, Miles Congreve¹, Robert M. Cooke¹ & Fiona H. Marshall¹

Chemokines and their G-protein-coupled receptors play a diverse role in immune defence by controlling the migration, activation and survival of immune cells¹. They are also involved in viral entry, tumour growth and metastasis and hence are important drug targets in a wide range of diseases^{2,3}. Despite very significant efforts by the pharmaceutical industry to develop drugs, with over 50 small-molecule drugs directed at the family entering clinical development, only two compounds have reached the market: maraviroc (CCR5) for HIV infection and plerixafor (CXCR4) for stem-cell mobilization⁴. The high failure rate may in part be due to limited understanding of the mechanism of action of chemokine antagonists and an inability to optimize compounds in the absence of structural information⁵. CC chemokine receptor type 9 (CCR9) activation by CCL25 plays a key role in leukocyte recruitment to the gut and represents a therapeutic target in inflammatory bowel disease⁶. The selective CCR9 antagonist vercirnon progressed to phase 3 clinical trials in Crohn's disease but efficacy was limited, with the need for very high doses to block receptor activation⁶. Here we report the crystal structure of the CCR9 receptor in complex with vercirnon at 2.8 Å resolution. Remarkably, vercirnon binds to the intracellular side of the receptor, exerting allosteric antagonism and preventing G-protein coupling. This binding site explains the need for relatively lipophilic ligands and describes another example of an allosteric site on G-protein-coupled receptors⁷ that can be targeted for drug design, not only at CCR9, but potentially extending to other chemokine receptors.

To obtain a crystal structure of human CCR9, a thermostabilized receptor (StaR) was generated^{8,9} containing eight amino-acid

substitutions (Extended Data Figs 1 and 2). These modifications did not alter vercirnon binding properties of the receptor compared with wild-type (Extended Data Fig. 3); however, stabilization with the [³H] vercirnon antagonist precludes G-protein coupling of the final StaR (Data not shown). To further facilitate crystallization, amino (N) and carboxy (C) termini were truncated resulting in the construct designated CCR9-StaR(25–340). No fusion partner(s) were used to aid crystallization, and the receptor was crystallized in lipidic cubic phase (LCP) in the presence of the antagonist vercirnon¹⁰ (4-*tert*-butyl-N-{4-chloro-2-[(1-oxidopyridin-4-yl)carbonyl]phenyl}benzenesulfonamide, GSK1605786, CCX282-B). The structure was determined to 2.8 Å resolution with two copies in the asymmetric unit arranged in a parallel fashion with TM4–TM4-mediated interactions (Extended Data Fig. 4). Details of data collection and refinement are in Extended Data Table 1. For discussion purposes, molecule A is used forthwith.

CCR9 exhibits the core canonical arrangement of seven transmembrane helices (TM1–TM7) with continuous density observed for all intracellular loops (ICLs) and helix 8 (Fig. 1a). Only extracellular loop 3 was resolved on the extracellular side of the receptor. Additionally, only residual signal is present for the conserved disulfide bridging the top of TM3 (Cys119^{3,25}) and extracellular loop 2. A second disulfide is present in CCR9 linking the N terminus (Cys38) with the top of TM7 (Cys289^{7,25}) as for the related chemokine receptor structures of CCR5/maraviroc¹¹ and CXCR4/IT1t¹². Structural superposition of the 7TM core of CCR9 with both CCR5 and CXCR4 (sequence identity 35%, Extended Data Fig. 5) achieves a Cα root mean square deviation of 1.9 Å and 2.5 Å, respectively, with the main differences across the extracellular halves of the receptors (Fig. 1b–g). Compared with CCR5 and

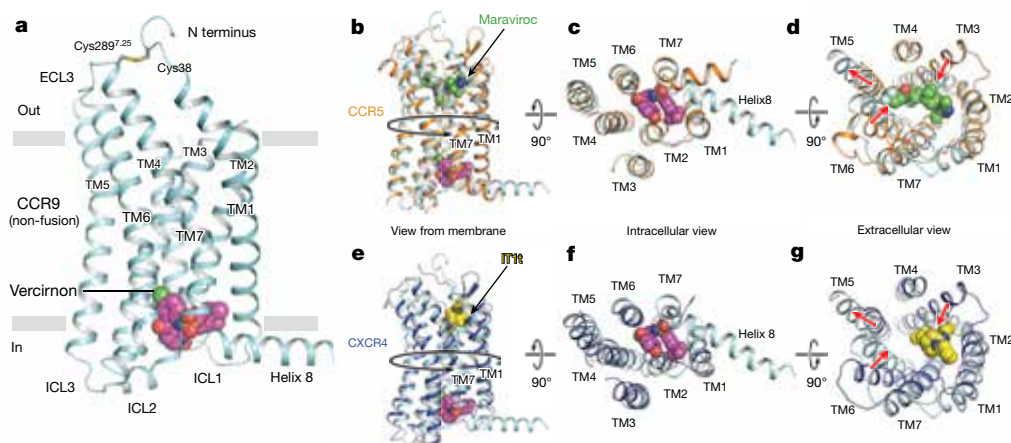


Figure 1 | Structure of CCR9 and comparison with CCR5 and CXCR4. **a**, Ribbon representation of CCR9 (cyan) viewed parallel to the membrane. Vercirnon is shown in sphere and stick representation, with carbon, nitrogen, chlorine, sulfur and oxygen atoms coloured magenta, blue, green, yellow and red, respectively. **b–d**, Superposition of CCR9 with

CCR5/maraviroc (orange, maraviroc in green) viewed from the membrane, intracellular and extracellular space, respectively. **e–g**, Superposition of CCR9 with CXCR4/IT1t (blue, IT1t in yellow) viewed from the membrane, intracellular and extracellular space, respectively. Significant changes in transmembrane positions are denoted by red arrows.

¹Heptares Therapeutics Ltd, BioPark, Broadwater Road, Welwyn Garden City, Hertfordshire AL7 3AX, UK.

*These authors contributed equally to this work.

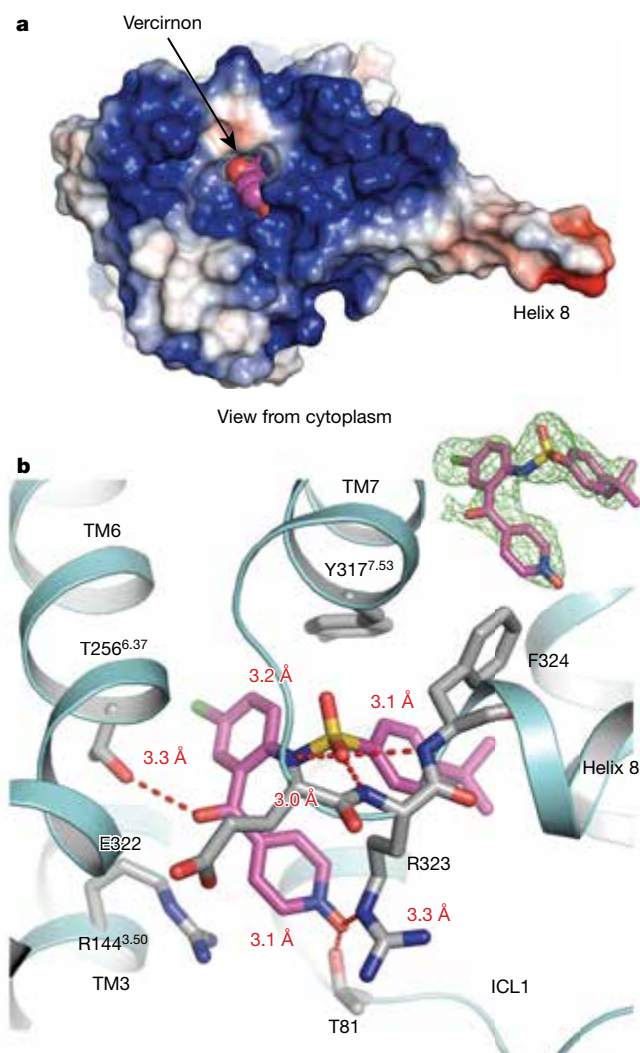


Figure 2 | Intracellular allosteric binding site of vercirnon in CCR9. **a**, Electrostatic surface representation of the intracellular surface of CCR9 with vercirnon bound (coloured as in Fig. 1) in the allosteric pocket open to the cytoplasm. **b**, Ligand interactions in the intracellular allosteric binding pocket; specific interactions are depicted as dashed red lines with distances labelled (inset) $F_0 - F_c$ OMIT density contoured at 2.0σ calculated before vercirnon inclusion in the model.

CXCR4, the tops of TM3 and TM6 of CCR9 are moved in towards the central axis of the helical bundle, and TM5 is moved outwards, with the differences being greatest between CCR9 and CXCR4. These changes in transmembrane helix position are possibly a consequence of the lack of a small molecule bound in the extracellular portion of the CCR9 transmembrane bundle.

Strong and unambiguous density is found for vercirnon on the intracellular side of the receptor contacting TM1, TM2, TM3, TM6, TM7 and helix 8 in an allosteric pocket within the helix bundle open to the cytoplasm (Fig. 2a and Extended Data Fig. 6). So far, the only other structural examples of small molecules binding towards the intracellular side of a receptor to effect allosteric antagonism are provided by the class B structures of corticotropin-releasing factor receptor type 1 (CRF₁R) in complex with the small-molecule antagonist CP-376395 (ref. 13) and the glucagon receptor (GCGR) in complex with MK-0893 (ref. 14). However, while CP-376395 is found in a pocket approximately 18 Å from the centre of the orthosteric cavity of CRF₁R, and MK-0893 adopts an extra-helical binding mode towards the bottom of TM6 in GCGR, the position of vercirnon bound to CCR9 is unique in both distance from the orthosteric site (approximately 33 Å) and in occupying a pocket with cytoplasmic access.

Moving to the molecular details of the CCR9–StAR–vercirnon interaction, the sulfone group of vercirnon hydrogen bonds with the backbone amino groups of Glu322, Arg323 and Phe324, acting as a helix cap for the N terminus of helix 8 in CCR9. Favourable interactions are also made with the side chain of Tyr317^{7.53} (of the conserved NP^{7.50}_{xxY(x)}_{5,6}F motif) from above the sulfone group. Mutation of Tyr317^{7.53}, Phe324 and Gly321^{7.57} to Ala, three highly conserved residues across all chemokine receptors (Fig. 3a, b and Extended Data Table 2), severely decreases vercirnon binding to CCR9 (Fig. 3c and Extended Data Fig. 7), highlighting the importance of these residues in forming the core scaffold of the intracellular allosteric binding site, with Gly321^{7.57} contributing the necessary conformational flexibility in the junction of TM7–helix-8 to orient the N terminus of helix 8 for ligand interaction.

The ligand pyridine-*N*-oxide group is oriented towards the intracellular face of the receptor at the cytoplasmic entrance to the ligand binding cavity. The pyridine-*N*-oxide is surrounded by polar residues located on the intracellular extremities of TM2, TM3 and the TM7–helix-8 hinge region including Thr83^{2.39}, Asp84^{2.40}, Arg144^{3.50}, Arg323 (on helix 8) Thr81 (on ICL1)—the last two being within hydrogen bonding distance of the pyridine-*N*-oxide (Fig. 2b). Mutation of Thr81 to glutamic acid reduces vercirnon binding compared with wild type (Fig. 3c), presumably as a result of the glutamic-acid side chain no longer being poised to make a polar contact with the ligand and/or fully engaging with Arg323 on helix 8. Finally, the ketone group of vercirnon is engaged in a hydrogen bond with the side chain of Thr256^{6.37},

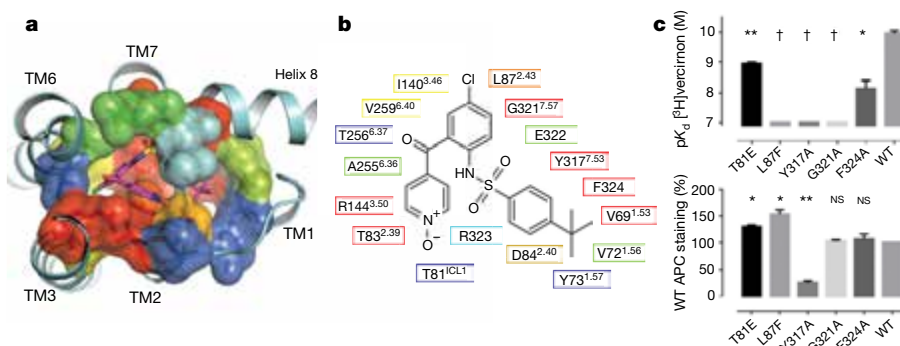


Figure 3 | Conservation and mutagenesis of vercirnon binding site. **a**, CCR9 allosteric site in surface representation with residues in rainbow spectrum according to conservation across chemokine receptors (red = 100%; blue = 0%). **b**, Two-dimensional schematic of **a**. c, [³H]vercirnon binding analysis of point mutations in allosteric site. Top: pK_d from saturation binding analysis; T81E ($P = 0.0027$), F324A ($P = 0.0116$). Bottom: cell

surface expression (percentage of wild-type (WT) allophycocyanin (APC) staining); T81E ($P = 0.0134$), L87F ($P = 0.0232$), Y317A ($P = 0.002$). Data shown as mean \pm s.e.m. representative of three independent experiments performed in duplicate. Statistical difference, represented with asterisks calculated from unpaired two-tailed *t*-tests. * $P \leq 0.05$; ** $P \leq 0.01$. †Ambiguous values due to near-complete loss of specific binding.

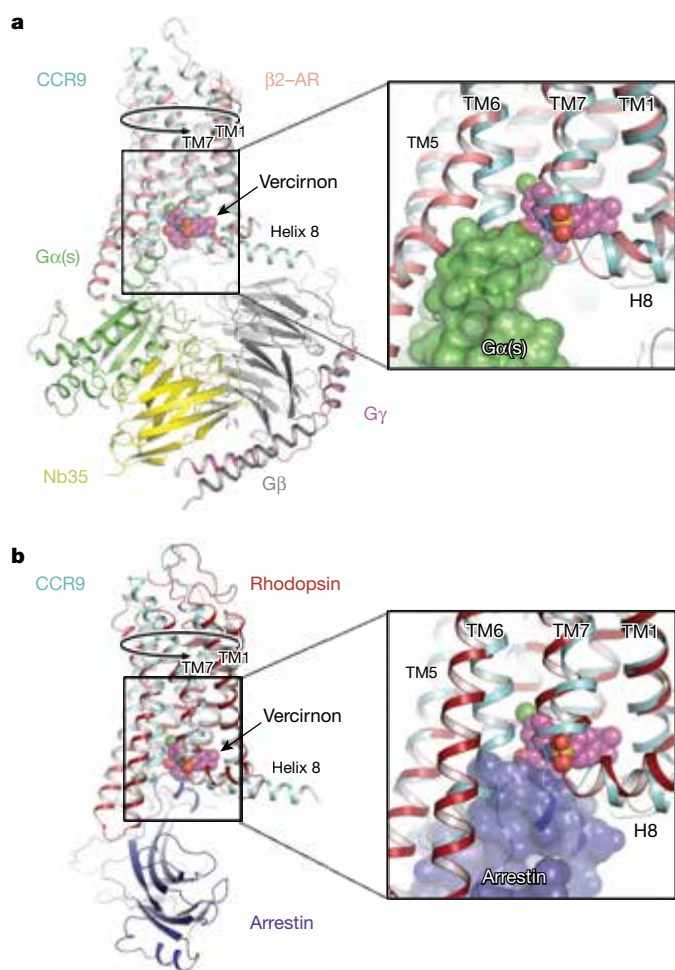


Figure 4 | Mechanism of vercirnon exerting intracellular antagonism of CCR9. **a**, Structural superposition of CCR9–vercirnon with the $\beta 2$ -AR–G_s complex structure (Protein Data Bank accession number 3SN6). **b**, Structural superposition of CCR9–vercirnon with the rhodopsin–arrestin complex structure (Protein Data Bank accession number 4ZWJ). Vercirnon exerts intracellular antagonism by holding the intracellular half of the receptor in a conformation sterically incompatible with G-protein or arrestin binding, both of which clash with the small molecule itself.

resulting in a ligand-mediated polar network linking TM6 across to ICL1, TM7 and the junction with helix 8 (Fig. 2b).

The *tert*-butylphenyl group anchors vercirnon in a cavity formed by TM1, TM2, TM7 and helix 8 and is characterized by concurrent lipophilic and hydrophilic residues. The lipophilic *tert*-butyl group faces towards the TM1–TM2 interface and makes hydrophobic interactions with Val69^{1.53}, Val72^{1.56}, Tyr73^{1.57} in TM1 and Leu87^{2.43} in TM2. Phe324 (on helix 8) and Tyr317^{7.53} make edge-to-face π – π stacking with the aromatic core of the *tert*-butylphenyl group (Fig. 2b). Other favourable hydrophobic interactions occur with the aliphatic portion of Arg323 and the side chain of Leu87^{2.43}, which is concomitantly engaged with the chlorophenyl part of the ligand.

The chlorophenyl moiety of vercirnon is located in a narrow, apolar cavity surrounded by several hydrophobic residues from TM2, TM3, TM6 and TM7. The chloro group, pointing up towards the central core of the receptor between TM3 and TM6, is located between the residues Leu87^{2.43}, Ile140^{3.46} and Val259^{6.40}. The aromatic part of the chlorophenyl group is located between the hydrophobic surface of Leu87^{2.43} and the main chain of Ala255^{6.36}. Mutation of Leu87^{2.43} to phenylalanine abolishes vercirnon binding to CCR9 (Fig. 3c), probably as a result of filling the cavity between TM2, TM3 and TM6 with a bulky aromatic side chain. The aromatic ring of Tyr317^{7.53} and the methyl

group of Thr83^{2.39} also contribute to favourable interactions with the chlorophenyl ring. The conservative stabilizing mutation V255A is found in the proximity of the chlorophenyl group of vercirnon; however, none of the stabilizing mutations altered vercirnon binding properties of the receptor compared with wild type (see earlier).

Molecular dynamics simulations of vercirnon bound to CCR9 (both StaR and wild type) showed stable interactions between the ligand and the residues in the binding site, with hydrogen bonds anchoring the sulfone group to the backbone of Arg323 and Phe324. However, after removal of ligand, molecular dynamics simulations of the pseudo-apo model showed a reorientation of side chains of Tyr317^{7.53}, Arg323 and Phe324 towards the centre of the transmembrane bundle (Extended Data Fig. 8). Interestingly, the corresponding region in the CCR5 (ref. 11) structure is similar to the CCR9 pseudo-apo model after molecular dynamics.

Small-molecule chemokine receptor antagonists may be split into two broad chemical classes: tertiary amines and non-amines. Tertiary amines represent most compounds identified so far and probably engage a buried acidic residue (E283^{7.39} in the CCR5–maraviroc complex¹¹) in the now well-understood class A transmembrane ligand-binding site region, explaining the preponderance of these molecules in chemical literature. Non-amines, such as vercirnon, have been less frequently reported and display pharmacological properties inconsistent with typical receptor antagonism. Interestingly, pepducin ATI-2341, a potent agonist of CXCR-type receptor 4 (CXCR4) and whose peptide sequence derives from the first ICL of the receptor, suggests modulation of receptor activity by acting at the intracellular receptor surface¹⁵. Furthermore, mutagenesis studies have repeatedly suggested that many of the non-amine class of chemokine antagonists bind near the intracellular surface of receptors, for example the highly CCR4 selective pyrazinyl-sulfonamide series¹⁶. For the dual CXCR1/2 squaramide antagonist SCH-527123, mutagenesis of CXCR2 suggests an intracellular allosteric pocket^{17,18} lined by Thr83^{2.39}, Asp84^{2.40}, Tyr314^{7.53} and Lys320^{7.59}, correlating with the vercirnon binding site in CCR9 (Extended Data Fig. 5); indeed a similar intracellular interaction mode may also exist for SB-656933 (ref. 19) binding to CXCR2. Additionally, investigation of two CXCR2 antagonists exhibiting 1000-fold selectivity over CXCR1, shows that selectivity can be reversed by swapping the receptor C-terminal tails, specifically mapping to residue Lys/Asn^{7.59} (correlating to Arg323^{7.59} in CCR9 which makes a direct contact to the pyridine-*N*-oxide of vercirnon)²⁰. Pharmacological evidence for an intracellular allosteric binding site in CXCR2 is further provided by the insurmountable inhibition of CXCL8-promoted β -arrestin-2 recruitment by SB-265610 (ref. 21). Triazolopyridylbenzenesulfonamides (CCR2), indazolesulfonamides (CCR4), repertaxin (CXCR1) and dihydroquinazolines (CXCR3) represent additional examples of non-amine chemokine antagonists⁵. The chemical similarity of several of these compounds to vercirnon, particularly the CCR2 and CCR4 antagonists that contain an aromatic sulfonamide (found capping helix 8 in CCR9), is highly suggestive of analogous sites on the intracellular face of their respective receptors. Overall, a consideration of the chemical nature of non-amine ligand classes, their pharmacological behaviour and evidence from mutagenesis supports the notion that an intracellular binding site may exist in many chemokine receptors, and that subtype-selective ligands can often be identified. Resolution of the structural details of this site in CCR9 facilitates further studies of non-amine chemokine antagonists using structure-based drug design.

In response to chemokine binding, CCR9 and chemokine receptor signalling in general have been most widely characterized via the heterotrimeric G-protein G_i transducer. However, C-terminal receptor phosphorylation by GRK can mediate β -arrestin binding, desensitization and internalization, alongside activation of, for example, Src, PI3K and MAPK²², with vercirnon inhibiting such signalling¹⁰. In structural terms, class A receptor agonist binding elicits a rigid-body movement along TM6, altering the interface to TM5 and causing an outward movement of the intracellular half of TM6 alongside an upward

movement of TM3 (refs 23, 24). Superposition of CCR9–vercirnon with the β_2 –AR–G_s complex structure²⁵ using the core transmembrane bundles provides a structural basis for intracellular allosteric antagonism (Fig. 4a). Assuming that G_i binds analogously, the G-protein clashes with vercirnon and TM6/ICL3 of CCR9, a likely consequence of vercirnon mediating a network of polar contacts (see earlier) from TM6 across to TM7/helix 8 and ICL1, which holds TM6 inwards towards the receptor's central helical axis. This, alongside acting as a steric wedge within the helical bundle, restricts the required movements of TM6/TM3, thereby abrogating G-protein binding. Superposition with the structure of rhodopsin bound to arrestin²⁶ demonstrates a similar situation where vercirnon–CCR9 interactions specifically occupy two of the major arrestin–receptor interfaces. Additionally, the junction of TM7–helix-8 in rhodopsin and the finger loop of arrestin directly clash with vercirnon (Fig. 4b).

The structure of CCR9 complexed with vercirnon provides the first detailed view of a small molecule bound on the intracellular surface of a G-protein-coupled receptor, in a pocket within the helical bundle of the receptor but open to the cytoplasm. This novel allosteric pocket may be targeted for the design of selective small-molecule antagonists of CCR9 (or related chemokine receptors). Since the intracellular regions of the receptor that interact with G proteins are overlapping but not identical to those that engage β -arrestin, a unique opportunity may now exist to deploy structure-based drug design techniques in fine-tuning molecules that differentially modulate biased signalling cascades and functional outcomes in the chemokine receptor family.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 August; accepted 7 November 2016.

Published online 7 December 2016.

1. Pease, J. E. Targeting chemokine receptors in allergic disease. *Biochem. J.* **434**, 11–24 (2011).
2. Wilkin, T. J. & Gulick, R. M. CCR5 antagonism in HIV infection: current concepts and future opportunities. *Annu. Rev. Med.* **63**, 81–93 (2012).
3. Vela, M., Aris, M., Llorente, M., Garcia-Sanz, J. A. & Kremer, L. Chemokine receptor-specific antibodies in cancer immunotherapy: achievements and challenges. *Front. Immunol.* **6**, 12 (2015).
4. Solari, R., Pease, J. E. & Begg, M. "Chemokine receptors as therapeutic targets: why aren't there more drugs?". *Eur. J. Pharmacol.* **746**, 363–367 (2015).
5. Pease, J. & Horuk, R. Chemokine receptor antagonists. *J. Med. Chem.* **55**, 9363–9392 (2012).
6. Wendt, E. & Keshav, S. CCR9 antagonism: potential in the treatment of inflammatory bowel disease. *Clin. Exp. Gastroenterol.* **8**, 119–130 (2015).
7. Changeux, J. P. & Christopoulos, A. Allosteric modulation as a unifying mechanism for receptor function and regulation. *Cell* **166**, 1084–1102 (2016).
8. Serrano-Vega, M. J., Magnani, F., Shibata, Y. & Tate, C. G. Conformational thermostabilization of the β_1 -adrenergic receptor in a detergent-resistant form. *Proc. Natl Acad. Sci. USA* **105**, 877–882 (2008).
9. Robertson, N. *et al.* The properties of thermostabilised G protein-coupled receptors (StaRs) and their use in drug discovery. *Neuropharmacology* **60**, 36–44 (2011).
10. Walters, M. J. *et al.* Characterization of CCX282-B, an orally bioavailable antagonist of the CCR9 chemokine receptor, for treatment of inflammatory bowel disease. *J. Pharmacol. Exp. Ther.* **335**, 61–69 (2010).

11. Tan, Q. *et al.* Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* **341**, 1387–1390 (2013).
12. Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330**, 1066–1071 (2010).
13. Hollenstein, K. *et al.* Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature* **499**, 438–443 (2013).
14. Jazayeri, A. *et al.* Extra-helical binding site of a glucagon receptor antagonist. *Nature* **533**, 274–277 (2016).
15. Tchernychev, B. *et al.* Discovery of a CXCR4 agonist pepducin that mobilizes bone marrow hematopoietic cells. *Proc. Natl Acad. Sci. USA* **107**, 22255–22259 (2010).
16. Andrews, G., Jones, C. & Wreggett, K. A. An intracellular allosteric site for a specific class of antagonists of the CC chemokine G protein-coupled receptors CCR4 and CCR5. *Mol. Pharmacol.* **73**, 855–867 (2008).
17. Gonsiorek, W. *et al.* Pharmacological characterization of Sch527123, a potent allosteric CXCR1/CXCR2 antagonist. *J. Pharmacol. Exp. Ther.* **322**, 477–485 (2007).
18. Salchow, K. *et al.* A common intracellular allosteric binding site for antagonists of the CXCR2 receptor. *Br. J. Pharmacol.* **159**, 1429–1439 (2010).
19. Lazaar, A. L. *et al.* SB-656933, a novel CXCR2 selective antagonist, inhibits *ex vivo* neutrophil activation and ozone-induced airway inflammation in humans. *Br. J. Clin. Pharmacol.* **72**, 282–293 (2011).
20. Nicholls, D. J. *et al.* Identification of a putative intracellular allosteric antagonist binding-site in the CXCR2 chemokine receptors 1 and 2. *Mol. Pharmacol.* **74**, 1193–1202 (2008).
21. de Kruijff, P. *et al.* Nonpeptidergic allosteric antagonists differentially bind to the CXCR2 chemokine receptor. *J. Pharmacol. Exp. Ther.* **329**, 783–790 (2009).
22. Thelen, M. Dancing to the tune of chemokines. *Nature Immunol.* **2**, 129–134 (2001).
23. Deupi, X. & Standfuss, J. Structural insights into agonist-induced activation of G-protein-coupled receptors. *Curr. Opin. Struct. Biol.* **21**, 541–551 (2011).
24. Tehan, B. G., Bortolato, A., Blaney, F. E., Weir, M. P. & Mason, J. S. Unifying family A GPCR theories of activation. *Pharmacol. Ther.* **143**, 51–60 (2014).
25. Rasmussen, S. G. *et al.* Crystal structure of the β_2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549–555 (2011).
26. Kang, Y. *et al.* Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature* **523**, 561–567 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Axford, R. Owen and D. Sherrell at I24, Diamond Light Source, Oxford, UK, for technical support. We thank colleagues at Heptares Therapeutics for suggestions and comments, and G. Brown and S. Bucknell for assistance in radioligand preparation.

Author Contributions J.K. and A.J. devised and performed the conformational thermostabilization and mutagenesis of the receptor, characterized expression constructs and performed radioligand binding analysis of mutants. Computational analysis of the structure and modelling was performed by F.D. and J.S.M. A.S.D. established the platform/protocols for LCP crystallization and solved the structure. J.C.E. supported expression and scouted purification of the final StaR. M.R. designed and characterized all constructs, collected and processed X-ray diffraction data and solved the structure. C.O. optimized purification, performed LCP crystallization, harvested crystals, collected and processed X-ray diffraction data, and solved and refined the structure. K.B. performed and analysed the pharmacology data. J.A.C. and M.C. identified and sourced the chemical compound(s) used in the study. Project management was performed by J.A.C., R.M.C. and F.H.M. The manuscript was prepared by A.S.D., C.O., F.D., M.C. and F.H.M. All authors contributed to the final editing and approval of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.H.M. (fiona.marshall@heptares.com).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Preparation of [^3H]verciron and verciron. Pyridine intermediate **1** (4-*tert*-butyl-*N*-[4-chloro-2-(pyridin-4-ylcarbonyl)phenyl]benzenesulfonamide) was prepared according to published procedures²⁷. Intermediate **2** and the radioligand [^3H]verciron **3** were prepared by Quotient Bioresearch (see Supplementary Fig. 1). Briefly, intermediate **2** was prepared by reaction of a solution of intermediate **1** in DCM with tritium gas in the presence of (1,5-cyclooctadiene)(tricyclohexylphosphine) (pyridine)iridium(I) hexafluorophosphate (Crabtree's catalyst), followed by purification by high-performance liquid chromatography (HPLC). Subsequently, [^3H]verciron was prepared by *N*-oxidation according to published procedures for the intermediate **1** (ref. 27) and purified by preparative HPLC. Mass spectrometry of [^3H]verciron gave a spectrum which was consistent with verciron and implied the incorporation of on average between one and two tritium atoms per molecule, which was consistent with the preparation of deuterated intermediate **2** in an analogous catalytic isotope exchange reaction with deuterium gas. The radiochemical was determined to have a purity of 99.9% by HPLC and a specific activity of 35 Ci/mmol. Cold verciron was prepared from intermediate **1** by *N*-oxidation according to published procedures²⁷ and purified by preparative thin-layer chromatography.

StaR generation. Full-length human CCR9 (1–369) was used as background for the generation of the conformationally thermostabilized receptor using a mutagenesis approach described earlier⁹. Mutants were analysed for thermostability in the presence of the radioligand [^3H]verciron. The CCR9-StaR is the full-length receptor with eight thermostabilizing mutations.

Cell culture. HEK293T cells were purchased from the American Type Culture Collection and were cultured in DMEM supplemented with 10% (v/v) fetal bovine serum (FBS). Cells were transfected using GeneJuice (Merck Millipore) according to the manufacturer's instructions and harvested after 48 h.

Thermostability measurement. Transiently transfected HEK293T cells were incubated in 50 mM HEPES–NaOH pH 7.5, 150 mM NaCl, supplemented with cOmplete Protease Inhibitor Cocktail tablets (Roche), with 1% (w/v) *n*-dodecyl- β -D-maltopyranoside (DDM) or 1% (w/v) *n*-decyl- β -D-maltopyranoside (DM) at 4°C for 1 h. All subsequent steps were performed at 4°C. Samples were incubated with 250 nM [^3H]verciron for 1 h and crude lysates cleared by centrifugation at 16,000g for 15 min. Thermostability of the receptor was determined as previously described¹⁴. Thermal stability (T_m) is defined as the temperature at which 50% ligand binding is retained.

FACS analysis. HEK293T cells transiently expressing CCR9–enhanced green fluorescent protein (eGFP) constructs and mock-transfected cells were harvested 40 h post-transfection using non-enzymatic cell dissociation solution (Sigma–Aldrich). Cells were washed with FACS buffer (PBS, 0.1% sodium azide, supplemented with cOmplete Protease Inhibitor Cocktail tablets (Roche)) before counting. Half a million cells per staining sample were taken and re-suspended in 200 μl FACS buffer containing 2% BSA and Mouse anti-CCR9 (R&D systems, MAB179) at 5 $\mu\text{g}/\text{ml}$. After incubation for 1 h at room temperature, samples were washed three times with 200 μl FACS buffer, then resuspended in 200 μl FACS buffer containing 2% BSA and APC-conjugated Goat anti-Mouse IgG_{2A} (Southern Biotech, 1080-11S) at 0.5 $\mu\text{g}/\text{ml}$ and incubated at room temperature for 1 h in the dark. The cells were washed three times with 200 μl FACS buffer and finally resuspended in 200 μl FACS buffer before FACS analysis using BD FACSCantoII and FACSDiva software. Bound APC was detected using excitation wavelength (λ_{ex}) = 633 nm and emission wavelength (λ_{em}) = 660 nm.

Radioligand binding. For saturation binding experiments HEK293 membranes transiently expressing CCR9 (5 μg per well) or CCR9-StaR(1–369) (2.5 μg per well) were incubated with varying concentrations of [^3H]verciron (final assay concentration \approx 0–50 nM) in the presence or absence of 1 μM verciron to define non-specific binding (assay buffer: 25 mM HEPES–NaOH pH 7.1, 140 mM NaCl, 1 mM CaCl_2 , 5 mM MgCl_2 , 0.2% BSA). Binding assays were incubated for 3 h at 25°C. The reaction was terminated by rapid filtration through 96-well GF/B filter plates pre-soaked with 0.1% polyethyleneimine (PEI) using a 96-well head harvester (Tomtec, USA) and plates washed with 5×0.5 mL phosphate buffered saline. For saturation binding experiments of mutants, HEK293T cells transiently expressing CCR9–eGFP constructs or mock transfected cells were resuspended in buffer (50 mM HEPES–NaOH pH 7.5, 150 mM NaCl, supplemented with cOmplete Protease Inhibitor Cocktail tablets (Roche)) and homogenized using a Tissuemiser. Homogenized cells (5×10^4 cells per well) were incubated with varying concentrations of [^3H]verciron (final assay concentration \approx 0–15 nM) for 2.5 h at 25°C. Non-specific binding was defined using mock transfected cells. The reaction was terminated by rapid filtration through 96-well GF/C filter plates pre-soaked with Milli-Q water using a 96-well head harvester (Tomtec, USA) and plates washed

with 5×1 mL Milli-Q water. Specific binding was determined by subtracting mock transfected controls. Plates were dried, and bound radioactivity was measured using scintillation spectroscopy on a Microbeta counter (PerkinElmer, UK). Data were analysed using GraphPad Prism version 5 (San Diego, USA). Saturation binding data was globally fitted to one site total and non-specific binding, or one site-specific binding.

Truncation constructs. A panel of N- and C-terminal truncation variants of CCR9 was designed on the basis of multiple sequence alignment of all human chemokine receptors and secondary structure prediction^{28,29}. Truncated receptors were expressed in HEK293T cells as C-terminal fusions with eGFP followed by a deca-histidine tag. Receptors were solubilized in 50 mM HEPES–NaOH pH 7.5, 150 mM NaCl, and 1% (w/v) *n*-dodecyl- β -D-maltopyranoside (DDM) and 0.05% (v/v) cholesteryl hemisuccinate (CHS) and their expression levels and stability was assayed by whole-cell fluorescence, western-blotting and fluorescence-detection size-exclusion chromatography (fSEC) as described³⁰. The most suitable construct emerging from this screen comprised residues 25–340. Removal of post-translational modifications (glycosylation at Asn32 and putative palmitoylation at Cys337) was achieved by mutating residues Thr34—part of the glycosylation recognition sequence NXS/T—to Glu and Cys337 to Ala. Inclusion of an N-terminal GP64 signal sequence increased expression levels.

Expression, membrane preparation and protein purification. The truncated CCR9-StaR(25–340) construct was expressed with a C-terminal deca-histidine tag in *Spodoptera frugiperda* Sf21 cells (Oxford Expression Technologies) using ESF 921 medium (Expression Systems) supplemented with 10% (v/v) fetal bovine serum (Sigma–Aldrich) and 1% (v/v) penicillin/streptomycin (PAA Laboratories) with a Bac to Bac Expression System (Invitrogen). Cells were infected at a density of 2×10^6 to 3×10^6 cells per millilitre with baculovirus at an approximate multiplicity of infection of 1. Cultures were grown at 27°C with constant shaking and harvested by centrifugation 72 h after infection.

All subsequent steps were performed at 4°C unless otherwise stated. Membranes were prepared by resuspension of cells in PBS supplemented with cOmplete Protease Inhibitor Cocktail tablets (Roche), 10 mM magnesium chloride and 5 $\mu\text{g}/\text{ml}$ DNaseI (Roche) followed by disruption using a microfluidizer at 60,000 pounds per square inch (M-110L Pneumatic, Microfluidics). Membranes were collected by ultracentrifugation at 204,700g, resuspended in 50 mM HEPES–NaOH pH 7.5, 250 mM NaCl with cOmplete Protease Inhibitor Cocktail tablets (Roche), and stored at -80°C until use.

To purify the receptor, membranes were thawed at room temperature and incubated with 10 μM verciron for 30 min before solubilization with 1.5% (w/v) *n*-decyl- β -D-maltopyranoside (DM) for 1 h. Insoluble material was removed by ultracentrifugation at 204,700g and the receptors were immobilized by batch binding to 2.5 ml of NiNTA resin (Qiagen). The resin was packed into an Omnitag column (Kinesis) and washed with ten column volumes of 20 mM HEPES–NaOH pH 7.5, 250 mM NaCl, 0.15% (w/v) *n*-decyl- β -D-maltopyranoside DM, and 10 μM verciron then for ten column volumes with the same buffer supplemented with 64 mM imidazole before bound material was eluted in buffer containing 400 mM imidazole. The protein was then concentrated using an Amicon Ultra-15 centrifugal concentrator (MerckMillipore), MWCO 50 kDa, and subjected to preparative SEC in 20 mM HEPES–NaOH pH 7.5, 150 mM NaCl, 0.15% (w/v) *n*-decyl- β -D-maltopyranoside (DM), and 10 μM verciron on a Superdex 200 10/300 Increase column (GE Healthcare). Receptor purity was analysed by SDS–polyacrylamide gel electrophoresis and liquid chromatography–mass spectrometry, and receptor monodispersity was assayed by analytical SEC. Fractions containing the pure, monomeric receptor were concentrated to 10–20 mg/ml in a Vivaspin 500 centrifugal concentrator (Sartorius). Protein concentration was determined using the receptor's calculated extinction coefficient at 280 nm ($\epsilon_{280,\text{calc}} = 56,225 \text{ M}^{-1} \text{ cm}^{-1}$) and confirmed by quantitative amino-acid analysis.

Crystallization. CCR9-StaR(25–340) was crystallized in LCP at 20°C. The protein was concentrated to ~ 16 mg/ml and mixed with monoolein (Nu-Check) supplemented with 10% (w/w) cholesterol (Sigma Aldrich) and 10 μM verciron using the twin-syringe method³¹. The final protein:lipid ratio was 40:60 (w/w). Boli (70 nl) were dispensed on 96-well glass bases and overlaid with 800 nl precipitant solution using a Mosquito LCP from TTP Labtech. Rod-shaped crystals (40–80 μm) of CCR9-StaR(25–340) were grown in 100 mM 2-(Bis(2-hydroxyethyl)amino)acetic acid (BICINE) at a pH range of 7.9–8.0, 200 mM sodium malonate, 28–43% (v/v) polyethylene glycol 400, 10 mM ammonium formate/ammonium nitrate/magnesium formate and 10 μM verciron. Single crystals were mounted for data collection and cryo-cooled in liquid nitrogen without the addition of further cryoprotectant. A complete dataset to 2.8 Å was obtained by merging diffraction data from ten crystals belonging to the triclinic space group P1.

Diffraction data collection and processing. X-ray diffraction data were measured on a Pilatus3 6M detector at Diamond Light Source beamline I24 using a

beam size of $6 \times 8 \mu\text{m}$ diameter. Crystals displayed diffraction initially out to 2.7 \AA after exposure to a beam attenuated down to 60% for 0.12 s per degree of oscillation. It was possible to collect approximately 25° of useful data from each crystal before radiation damage became severe. Further attenuation down to 30% of beam allowed collection of about 60° of useful data. Data from individual crystals were integrated using XDS³². Data merging and scaling was performed using the program AIMLESS from the CCP4 suite^{33,34}. Data collection statistics are reported in Extended Data Table 1.

Structure solution and refinement. The structure of CCR9-StaR(25–340) was solved by molecular replacement with the program Phaser³⁵ using truncated CCR5 (Protein Data Bank accession number 4MBS) as the search model looking for two copies. Here the fusion protein rubredoxin was removed from the CCR5 structure. Manual model building was performed in COOT³⁶ using sigma-A-weighted $2m|F_o| - |DF_c|$, $m|F_o| - D|F_c|$ maps together with simulated-annealing and simple composite omit maps calculated using Phenix³⁷. Initial refinement was performed with REFMAC5 (ref. 38) using maximum-likelihood restrained refinement in combination with the jelly-body protocol. Further and final stages of refinement were performed with Phenix.refine³⁹ with positional, individual isotropic B-factor refinement and TLS. The final refinement statistics are presented in Extended Data Table 1. Coordinates and structure factors have been deposited in the Protein Data Bank under accession number 5LWE.

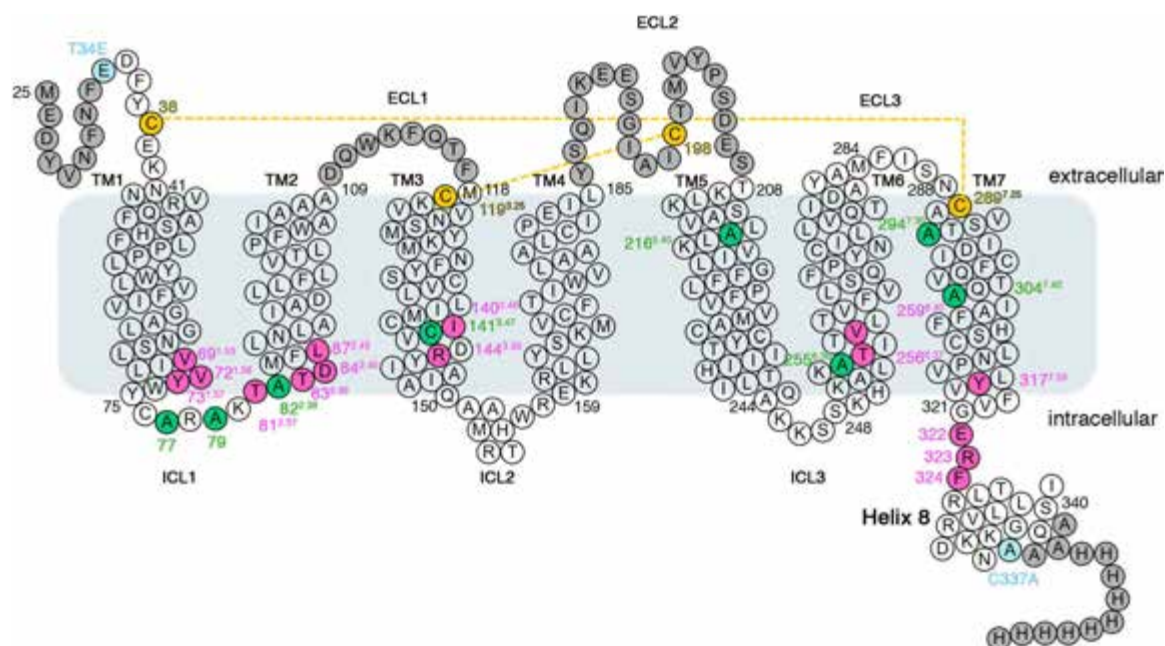
Structure analysis. Structures were superposed and aligned for comparison purposes using the program COOT³⁶ to generate global structural superpositions. Figures were prepared using PyMOL (Schrödinger, New York).

The CCR9/verciron structure was prepared with the Protein Preparation Wizard method in Maestro version 10.6 (Schrödinger, New York). Hydrogen atoms were energy minimized using the OPLS3 force field. The wild-type molecular model was created in Maestro by changing the StaR mutations to the correspondent wild-type residues. The system was embedded in an equilibrated POPC (1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine) bilayer and parameterized using the OPLS3 force field using the System Builder in Maestro. After the Relax protocol, the system was equilibrated for 100 ns molecular dynamics simulation using Desmond 4.6 (Desmond Molecular Dynamics System, D. E. Shaw Research, New York). The molecular dynamics was performed at 300K/1atm in the NPT ensemble using a Nose-Hoover thermostat and a Martyna-Tobias-Klein barostat⁴⁰

with a 2.0 ps relaxation time. Coulomb interactions were evaluated using a 9 \AA short-range cut-off and smooth particle mesh Ewald as long-range method (Ewald tolerance = 10^{-9}). The resulting molecular dynamics trajectories were analysed with the simulation interactions diagram method in Maestro.

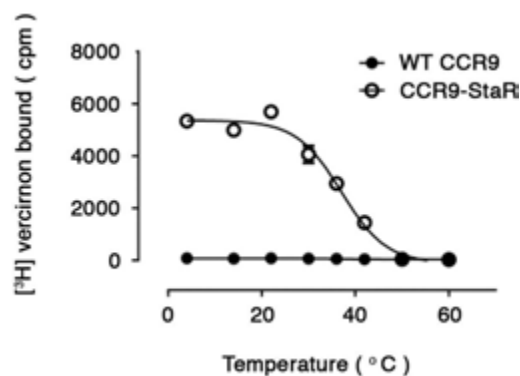
Data availability statement. Coordinates and structure factors have been deposited in the Protein Data Bank under the accession code 5LWE. All other data are available from the corresponding author upon reasonable request.

27. Ungashe, S. *et al.* Aryl sulphonamides. US patent 2006/0111351A1 (2006).
28. Nugent, T. & Jones, D. T. Membrane protein orientation and refinement using a knowledge-based statistical potential. *BMC Bioinformatics* **14**, 276–285 (2013).
29. Alva, V. *et al.* The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44** (Suppl. W1), W410–W415 (2016).
30. Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14**, 673–681 (2006).
31. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4**, 706–731 (2009).
32. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D* **66**, 133–144 (2010).
33. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
34. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69**, 1204–1214 (2013).
35. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
36. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
37. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
38. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
39. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
40. Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **101**, 4177–4189 (1994).
41. Diederichs, K. & Karplus, P. A. Better models by discarding data? *Acta Crystallogr. D* **69**, 1215–1222 (2013).



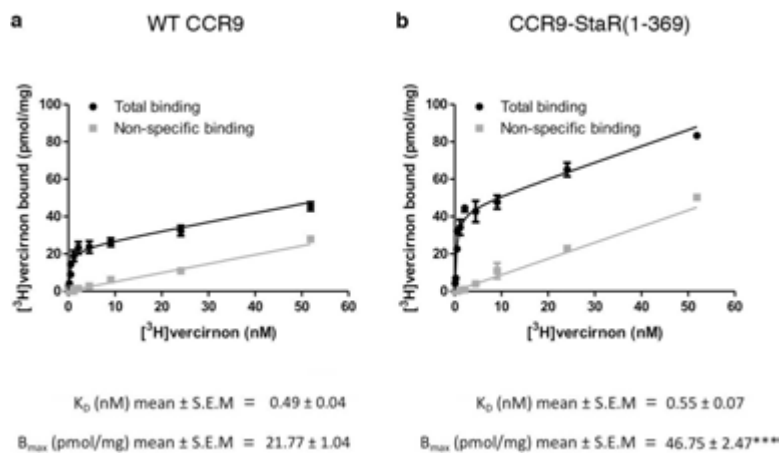
Extended Data Figure 1 | CCR9 crystallization construct StaR (25-340) in schematic representation. Thermostabilizing mutations (green) are Thr77Ala, Val79Ala, Met82Ala, Ser141Cys, Thr216Ala, Val255Ala, Asn294Ala, Thr304Ala. Further mutations to remove sites of post-translational modifications (light blue) are Cys337Ala and Thr34Glu.

Residues forming the allosteric pocket are pink. Disordered residues in the structure are grey. The disulfide bonds between (Cys119^{3,25}) and extracellular loop 2 and linking the N terminus (Cys38) with the top of TM7 (Cys289^{7,25}) are denoted by dashed yellow lines.



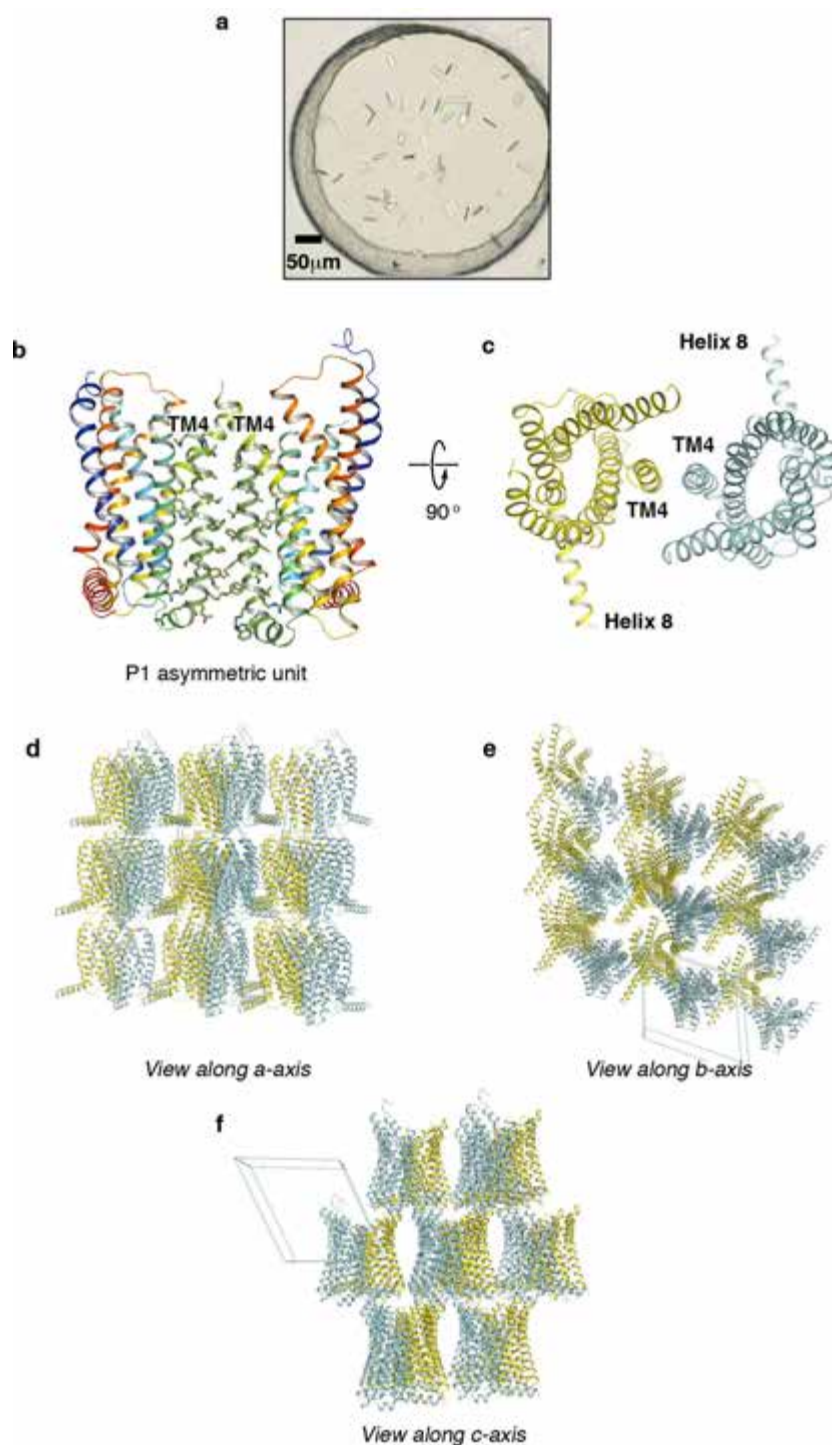
Extended Data Figure 2 | Comparison of wild-type and thermostabilized CCR9 in radioligand binding of [³H]verciron.

The thermal stability of wild-type CCR9 (filled circles) and CCR9-StaR(1-369) (open circles) analysed in decyl-maltoside are shown. Error bars are derived from standard deviations and calculated from duplicate temperature points ($n = 2$) within a single experiment. Data shown are representative of three independent experiments. CCR9-StaR(1-369) produced a mean T_m of 39.5°C. The T_m of wild-type CCR9 was not determinable under these conditions; however, binding is observed and a T_m can be calculated in dodecyl-maltoside (data not shown).



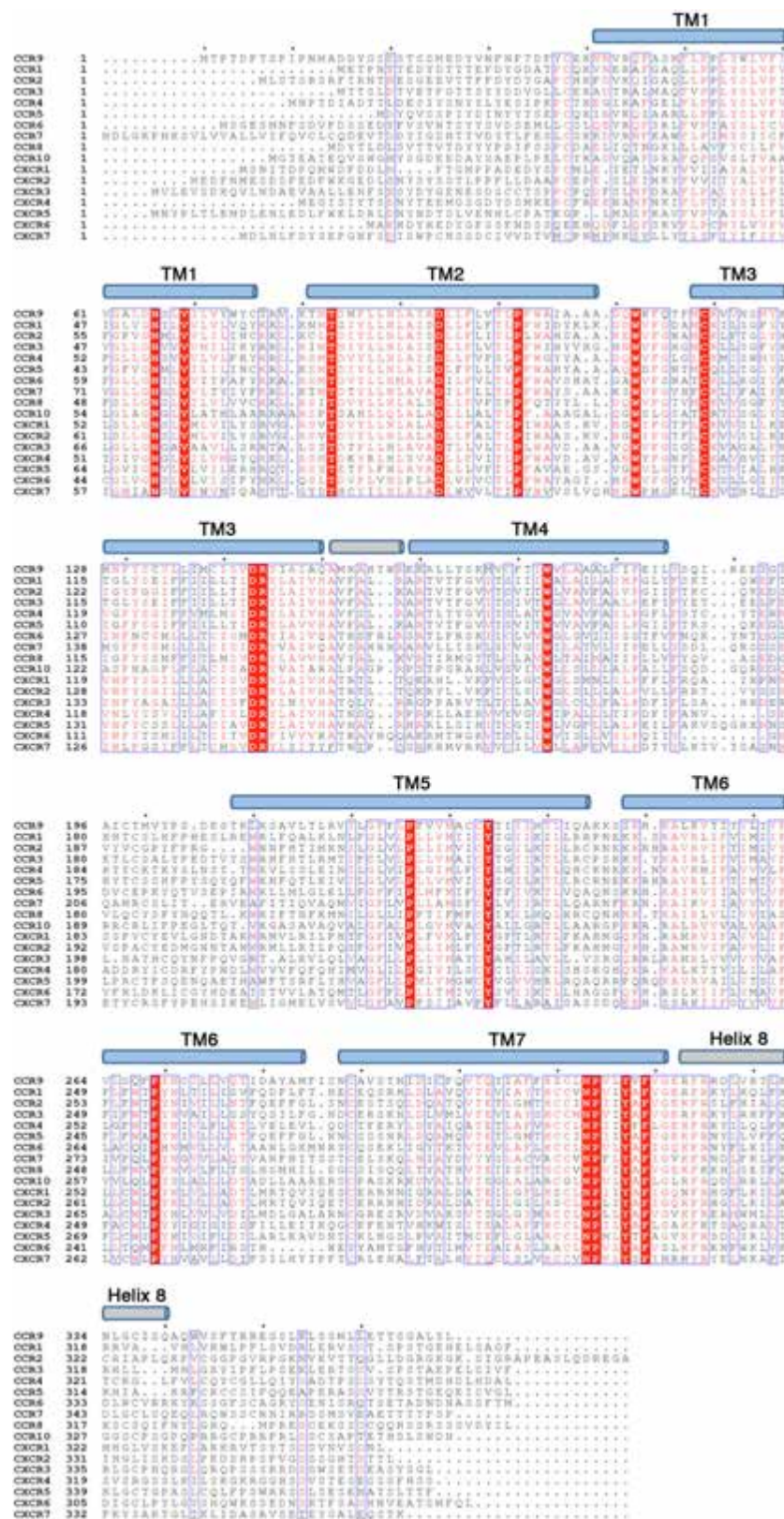
Extended Data Figure 3 | Pharmacology of WT CCR9 and CCR9-StaR. Saturation binding experiments performed in membranes from HEK293 cells transiently expressing (a) human CCR9 or (b) CCR9-StaR(1-369). Non-specific binding was determined by addition of 1 μ M cold vercirnon. Data shown as mean \pm s.e.m. are representative of three independent experiments performed in duplicate. Data were fitted globally to a one-site

saturation isotherm. Affinity and expression level (B_{max}) values are given below the graphs for both WT CCR9 and CCR9-StaR(1-369). There was no difference in the affinity of [3 H]vercirnon at WT CCR9 or CCR9-StaR(1-369) (unpaired, two-tailed t -test = 0.51). CCR9-StaR(1-369) showed significantly higher expression levels (B_{max}) than WT CCR9 (unpaired, two-tailed t -test = 0.0007).

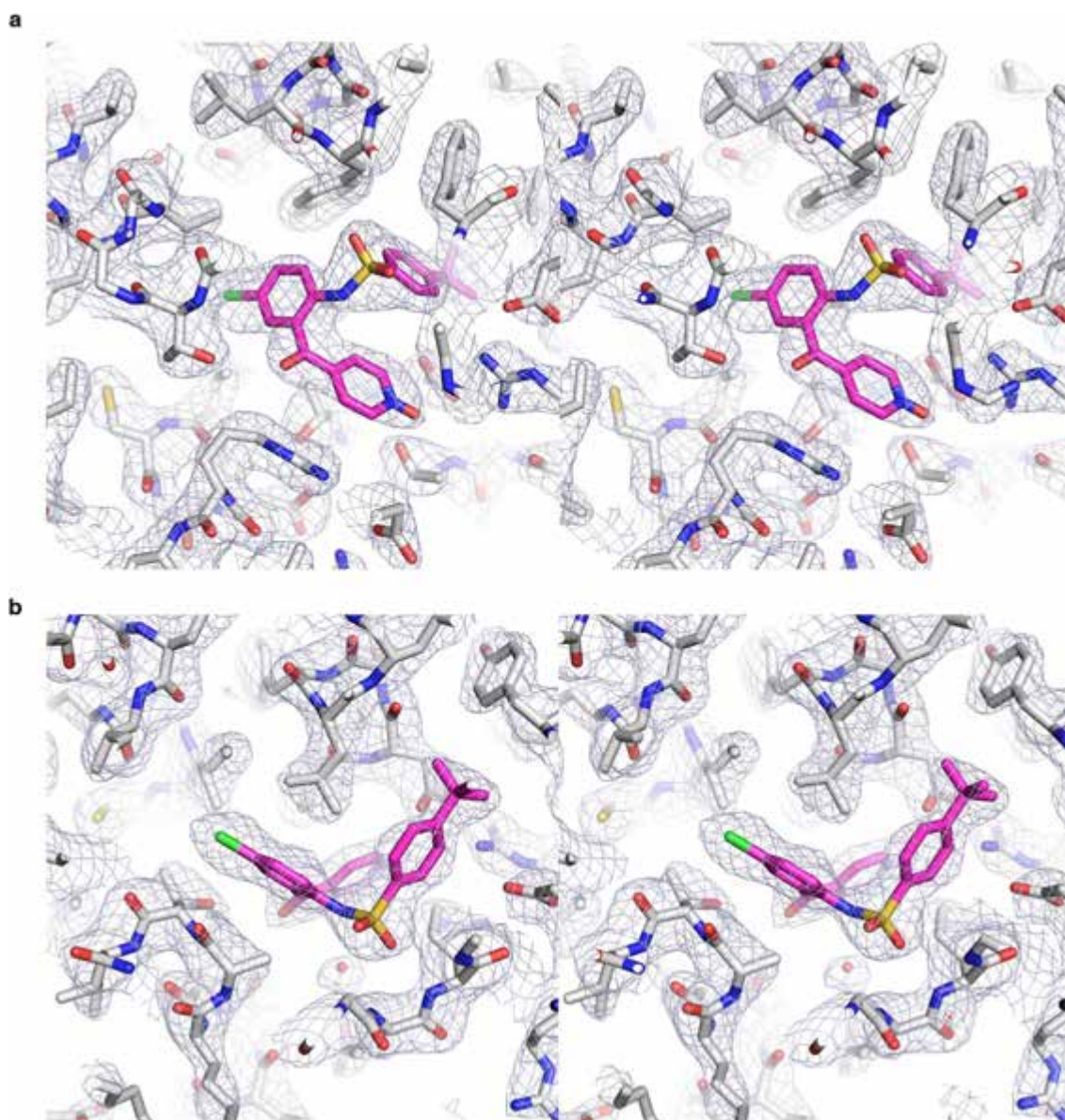


Extended Data Figure 4 | Crystal packing in the CCR9-StaR(25-340) triclinic system. **a**, Typical CCR9-StaR(25-340) non-fusion crystals grown in LCP complexed with vercirnon. **b**, The two copies of CCR9-StaR(25-340) in the triclinic asymmetric unit assemble in a parallel fashion with contacts mediated by TM4 – CCR9-StaR(25-340) shown in chainbow

colouration (blue to red equals N to C terminus). **c**, View as in **b** rotated by 90° with the two copies of CCR9-StaR(25-340) now coloured yellow and cyan. **d–f**, Views of CCR9-StaR(25-340) packing in the triclinic crystal system along the *a*, *b* and *c* axes respectively, molecules coloured as in **c**.

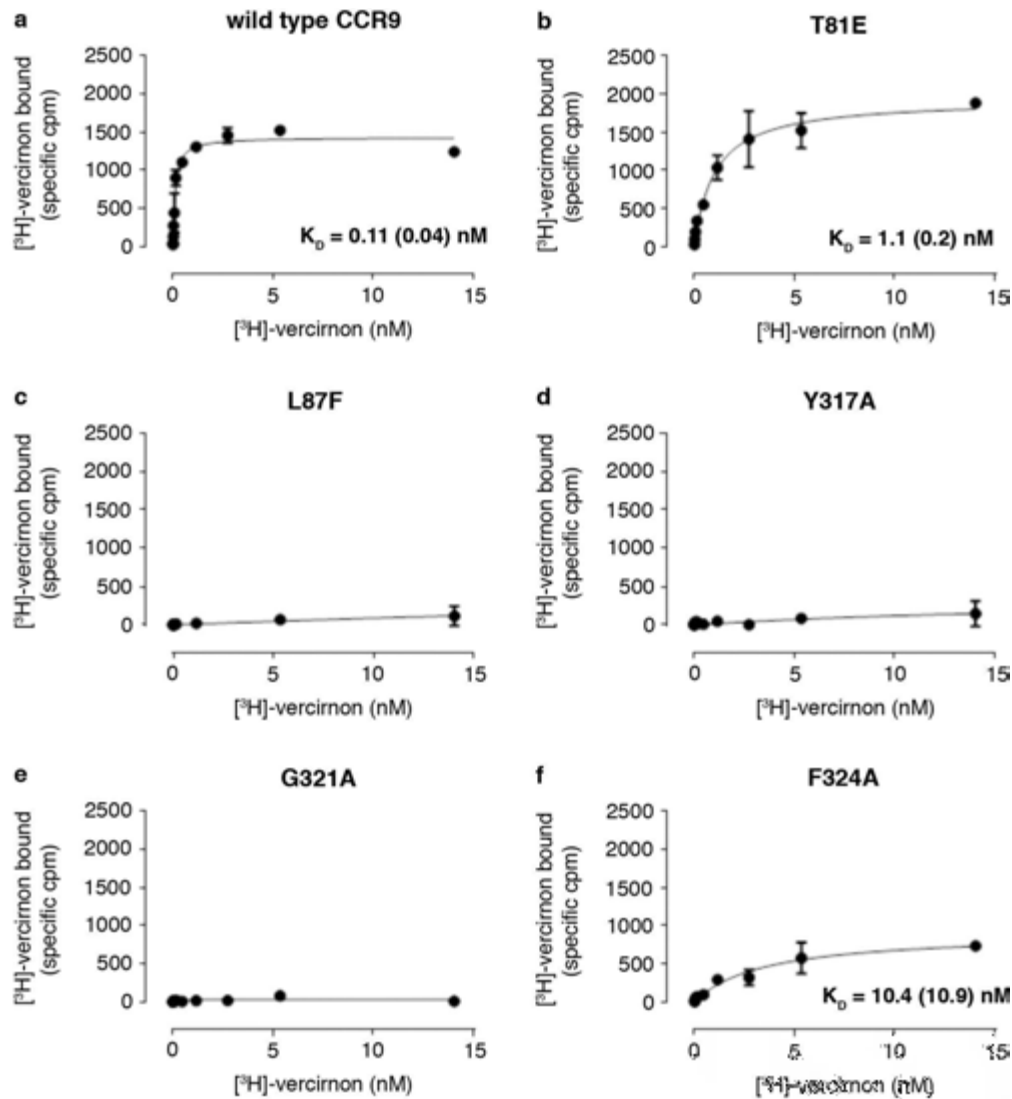


Extended Data Figure 5 | Multiple sequence alignment of human chemokine receptors.



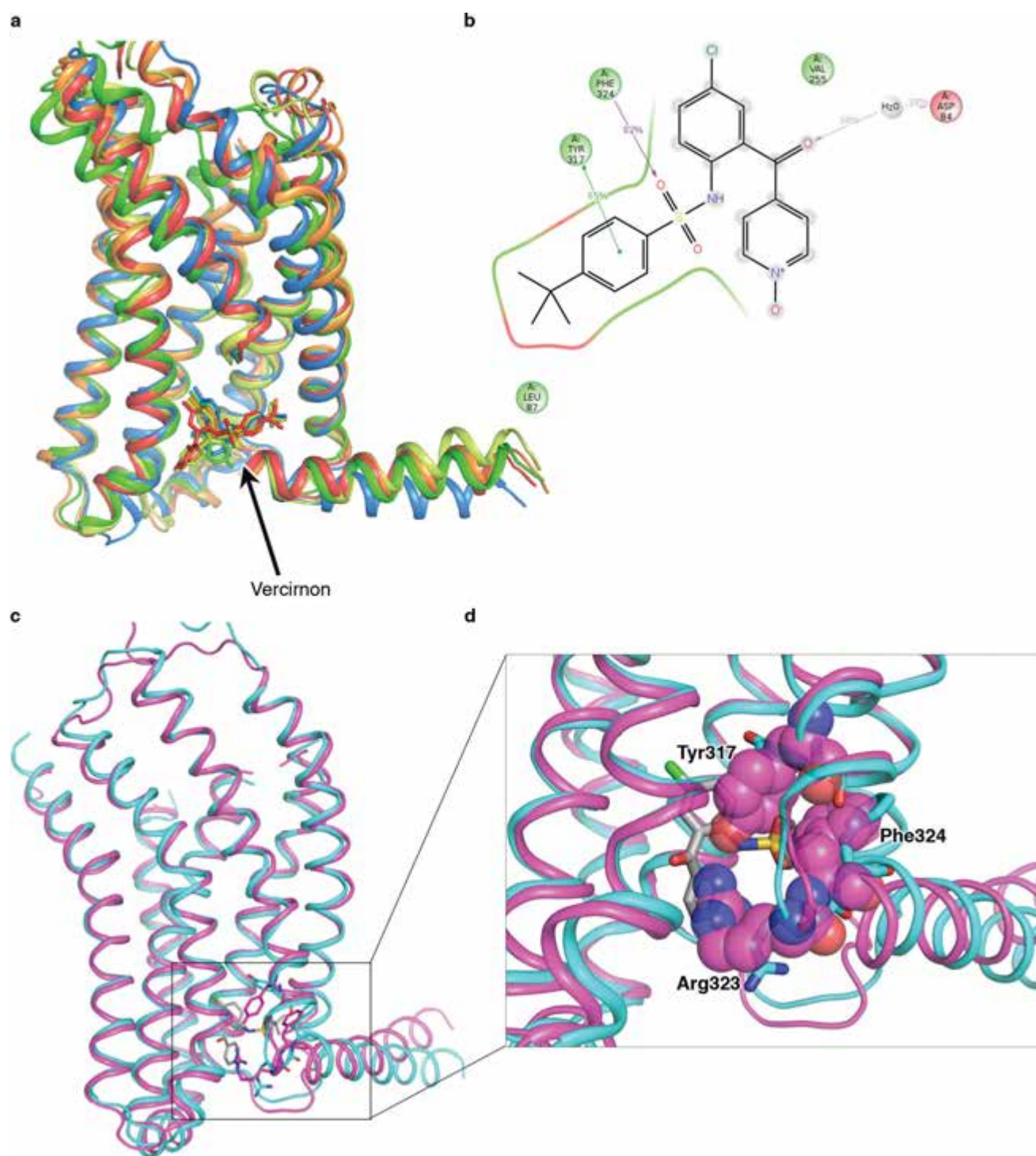
Extended Data Figure 6 | Electron density around the vercirnon binding site. a, Cross-eye stereoscopic view of $2F_o - F_c$ density contoured at 1.6σ covering vercirnon and surrounding residues as viewed from intracellular space. Vercirnon in stick representation, with carbon, nitrogen, chlorine,

sulfur and oxygen atoms coloured magenta, blue, green, yellow and red, respectively; CCR9 in stick representation with carbon, nitrogen, oxygen and sulfur atoms coloured white, blue, red and yellow respectively. **b,** View as in **a** rotated by 180° .



Extended Data Figure 7 | Saturation binding analysis of mutants with $[^3\text{H}]$ verciron. **a–f**, Saturation binding of $[^3\text{H}]$ verciron to homogenized cell lysates containing indicated mutant variants of CCR9. Data are representative of three independent experiments performed

in duplicate \pm s.d. K_D values (inset) are mean of three independent experiments with s.d. in parentheses. The datasets for L87F, Y317A and G321A could not be analysed unambiguously owing to near-complete loss of specific binding.



Extended Data Figure 8 | Molecular dynamics analysis of the CCR9-vercirnon complex. **a**, Stability of the CCR9-vercirnon complex during 100 ns molecular dynamics. Structural alignment of the wild-type CCR9-vercirnon complex at 0 (blue), 25 (green), 50 (yellow), 75 (orange) and 100 ns (red) molecular dynamics. Proteins are shown as ribbon with TM7 partly hidden for clarity; vercirnon is represented in sticks. **b**, Two-dimensional representation of the ligand-protein contacts. **c**, **d**, Induced-fit

binding of vercirnon to CCR9. Superposition of the CCR9-vercirnon complex (cyan) and the pseudo-apo state of CCR9 receptor at 100 ns molecular dynamics (magenta). Vercirnon is shown as sticks with carbons coloured in grey. Arg323, Phe324 and Tyr317 are shown as sticks with carbons coloured in cyan for the crystallographic structure and stick (**c**) or spheres (**d**) with carbons coloured in magenta for the molecular dynamics output.

Extended Data Table 1 | Data collection and refinement statistics for CCR9-StaR(25-340) complexed with vercirnon

Data collection	
Number of crystals	10
Space group	P1
Cell dimensions	
a, b, c (Å)	62.6, 66.2, 68.3
α , β , γ (°)	74.0, 64.7, 62.3
Number of reflections measured	78,953
Number of unique reflections	21,320
Resolution (Å)	58.34 - 2.80 (2.95 - 2.80)
R_{merge}	0.162 (0.887)
$CC_{1/2}$ **	0.980 (0.510)
Mean I/sd(I)	5.6 (1.7)
Completeness (%)	98.9 (98.3)
Redundancy	3.7 (3.7)
Refinement	
Resolution (Å)	19.97 - 2.80
Number of reflections (test set)	21,254 (1,133)
$R_{\text{work}}/R_{\text{free}}$	0.214 / 0.239
Number of atoms	
All	5,054
Protein	4,466
Ligand	60
Others (Lipids, ions, waters)	528
Average B factors (Å ²)	
All	69.1
CCR9	68.3
Ligand	41.9
Others (Lipid, ion, water)	79.1
RMSD	
Bond lengths (Å)	0.003
Bond angles (°)	0.552
Ramachandran statistics	
Favored regions (%)	99.3
Allowed regions (%)	0.7
Outliers (%)	0.0
<i>MolProbity</i> overall score (percentile)	1.35 (100th percentile)

*Values in parentheses indicate highest resolution shell. ** $CC_{1/2}$: see ref. 41.

Extended Data Table 2 | Conservation of vercirnon binding residues across all chemokine receptors

CCR9 (P51686)	Sulphone group					Pyridine- <i>N</i> -oxide group				
	Tyr317	Gly321	Glu322	Arg323	Phe324	Thr81	Thr83	Asp84	Arg144	Arg323
CCR1 (P32246)	Tyr	Gly	Glu	Arg	Phe	Asn	Thr	Ser	Arg	Arg
CCR2 (P41597)	Tyr	Gly	Glu	Lys	Phe	Cys	Thr	Asp	Arg	Lys
CCR3 (P51677)	Tyr	Gly	Glu	Arg	Phe	Ile	Thr	Asn	Arg	Arg
CCR4 (P51679)	Tyr	Gly	Glu	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CCR5 (P51681)	Tyr	Gly	Glu	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CCR6 (P51684)	Tyr	Gly	Gln	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CCR7 (P32248)	Tyr	Gly	Val	Lys	Phe	Thr	Thr	Asp	Arg	Lys
CCR8 (P51685)	Tyr	Gly	Glu	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CCR10 (P46092)	Tyr	Gly	Leu	Arg	Phe	Ser	Thr	Ser	Arg	Arg
CXCR1 (P25024)	Tyr	Gly	Gln	Asn	Phe	Ser	Thr	Asp	Arg	Asn
CXCR2 (P25025)	Tyr	Gly	Gln	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CXCR3 (P49682)	Tyr	Gly	Val	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CXCR4 (P61073)	Tyr	Gly	Ala	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CXCR5 (P32302)	Tyr	Gly	Val	Lys	Phe	Ser	Thr	Glu	Arg	Lys
CXCR6 (O00574)	Tyr	Ser	Leu	Lys	Phe	Ser	Thr	Asp	Arg	Lys
CXCR7 (P25106)	Tyr	Asn	Arg	Asn	Tyr	Tyr	Thr	His	Arg	Asn

CCR9 (P51686)	Tert-butyl-phenyl group							Chlorophenyl-ketone group			
	Val69	Val72	Tyr73	Leu87	Tyr317	Arg323	Phe324	Ile140	Val255	Thr256	Val259
CCR1 (P32246)	Val	Val	Leu	Leu	Tyr	Arg	Phe	Leu	Leu	Ile	Ile
CCR2 (P41597)	Val	Ile	Leu	Leu	Tyr	Lys	Phe	Leu	Val	Ile	Ile
CCR3 (P51677)	Val	Ile	Leu	Leu	Tyr	Arg	Phe	Leu	Leu	Ile	Ile
CCR4 (P51679)	Val	Val	Leu	Leu	Tyr	Lys	Phe	Met	Met	Ile	Val
CCR5 (P51681)	Val	Ile	Leu	Leu	Tyr	Lys	Phe	Leu	Leu	Ile	Ile
CCR6 (P51684)	Val	Thr	Phe	Leu	Tyr	Lys	Phe	Ile	Val	Ile	Val
CCR7 (P32248)	Val	Thr	Tyr	Leu	Tyr	Lys	Phe	Ile	Val	Ile	Val
CCR8 (P51685)	Val	Val	Leu	Leu	Tyr	Lys	Phe	Met	Leu	Val	Val
CCR10 (P46092)	Val	Thr	His	Leu	Tyr	Arg	Phe	Ile	Val	Val	Leu
CXCR1 (P25024)	Val	Val	Ile	Leu	Tyr	Asn	Phe	Ile	Val	Ile	Val
CXCR2 (P25025)	Val	Val	Ile	Leu	Tyr	Lys	Phe	Ile	Val	Ile	Val
CXCR3 (P49682)	Val	Val	Leu	Leu	Tyr	Lys	Phe	Ile	Leu	Val	Val
CXCR4 (P61073)	Val	Val	Met	Arg	Tyr	Lys	Phe	Ile	Thr	Thr	Leu
CXCR5 (P32302)	Val	Ile	Leu	Leu	Tyr	Lys	Phe	Ile	Val	Ala	Val
CXCR6 (O00574)	Val	Ile	Ser	Leu	Tyr	Lys	Phe	Ile	Ile	Ile	Val
CXCR7 (P25106)	Val	Val	Asn	Ile	Tyr	Asn	Phe	Met	Ile	Ile	Tyr

Structure and regulation of the chromatin remodeller ISWI

Lijuan Yan^{1,2*}, Li Wang^{1,2*}, Yuanyuan Tian^{1,2}, Xian Xia^{1,2} & Zhucheng Chen^{1,2}

ISWI is a member of the SWI2/SNF2 family of chromatin remodellers^{1,2}, which also includes Snf2, Chd1, and Ino80. ISWI is the catalytic subunit of several chromatin remodelling complexes, which mobilize nucleosomes along genomic DNA, promoting replication progression, transcription repression, heterochromatin formation, and many other nuclear processes^{3–5}. The ATPase motor of ISWI is an autonomous remodelling machine⁶, whereas its carboxy (C)-terminal HAND–SAND–SLIDE (HSS) domain functions in binding extranucleosomal linker DNA^{7–10}. The activity of the catalytic core of ISWI is inhibited by the regulatory AutoN and NegC domains, which are in turn antagonized by the H4 tail and extranucleosomal DNA, respectively, to ensure the appropriate chromatin landscape in cells¹¹. How AutoN and NegC inhibit ISWI and regulate its nucleosome-centring activity remains elusive. Here we report the crystal structures of ISWI from the thermophilic yeast *Myceliophthora thermophila* and its complex with a histone H4 peptide. Our data show the amino (N)-terminal AutoN domain contains two inhibitory elements, which collectively bind the second RecA-like domain (core2), holding the enzyme in an inactive conformation. The H4 peptide binds to the core2 domain coincident with one of the AutoN-binding sites, explaining the ISWI activation by H4. The H4-binding surface is conserved in Snf2 and functions beyond AutoN regulation. The C-terminal NegC domain is involved in binding to the core2 domain and functions as an allosteric element for ISWI to respond to the extranucleosomal DNA length.

We crystallized a construct of ISWI containing the catalytic core from *M. thermophila* (residues 81–723; Core; Fig. 1a), and refer to it here as MtISWI. The sequence of MtISWI is about 68%, 68%, and 58% identical to those of ISW1 and ISW2 of *Saccharomyces cerevisiae* and human SNF2h, respectively (Extended Data Fig. 1). The final structure was refined to 2.4 Å (Extended Data Table 1).

MtISWI folds into a compact structure, with the two RecA-like ATPase core domains (core1 and core2) packing together (Fig. 1b, c). We extend the concept of AutoN to include approximately 100 residues upstream the core1 domain¹¹. This newly defined AutoN domain binds the core1 domain through an N-terminal helical domain ($\alpha 1$ – $\alpha 3$), then interacts with the core2 domain via a long loop (L3) and a helix ($\alpha 4$), and finally connects back to the core1 domain. The NegC domain contains two helices ($\alpha 23$ and $\alpha 24$), extends out from the bulk of the protein, and interacts with a nearby molecule in the crystals (Fig. 1c and Extended Data Fig. 2a). Comparison with the structures of Chd1 and Snf2 indicates that although the individual RecA-like ATPase core domains share conserved folds and similar (but not identical) DNA-binding elements, these enzymes differ greatly in their overall organization (Extended Data Fig. 3)^{12,13}. The structural divergence provides the rational basis of the different regulation of these proteins.

The structure explains the nature of ISWI inhibition by AutoN. The two essential elements for ATP hydrolysis, the nucleotide-binding motif I (P-loop) and the catalytic motif VI (arginine fingers), were identified on the basis of sequence conservation (Extended Data Fig. 1).

Whereas the P-loop is exposed to solvent, motif VI is buried within the core1–core2 interface (Fig. 2a, b). Thus, the structure of ISWI is not compatible with efficient ATP hydrolysis and represents an inactive conformation. The two core domains of ISWI are mostly glued together by AutoN (Fig. 2a). Helices $\alpha 1$, $\alpha 2$ and the intervening loop at the N terminus of AutoN bind to the core1 domain through multiple hydrophobic residues (Fig. 2c), including Leu105 and Phe109, which are highly conserved in the ISWI subfamily of proteins, but not in the Chd1 or Snf2 remodellers, representing one of the characteristic features of ISWI (Extended Data Fig. 1).

AutoN bridges the core1 and core2 domains through $\alpha 3$, and then binds to the core2 domain mainly via H-bond/salt-bridge interactions through two elements, the L3 loop and the $\alpha 4$ helix (Fig. 2a). It is the L3 loop that contains the H4-like sequence¹¹. Arg151 of the L3 loop is involved in a pair of strong salt-bridge interactions with Asp524 of the core2 domain (Fig. 2d). Arg149 of the L3 loop H-bonds to Asp520. Consistent with the previous studies¹¹, mutations of the conserved Arg149 and Arg151 of MtISWI increased the ATPase activities of the protein (Fig. 2e). Besides these conserved interactions, we also found a pair of salt-bridge interactions between Glu474 and Arg141 (Fig. 2d). Arg141 is not found in the other ISWI proteins examined (Extended Data Fig. 1). Mutation of Arg141 also released the ISWI inhibition (Fig. 2e), suggesting Arg141 of MtISWI provides an additional degree of inhibition.

Multiple residues of the L3 loop seem to cooperatively bind the acidic surface of the core2 domain. Whereas ISWI with single point mutations (R141A, R149A, and R151A) showed comparable and slightly increased remodelling activities relative to the enzyme with intact interface, the

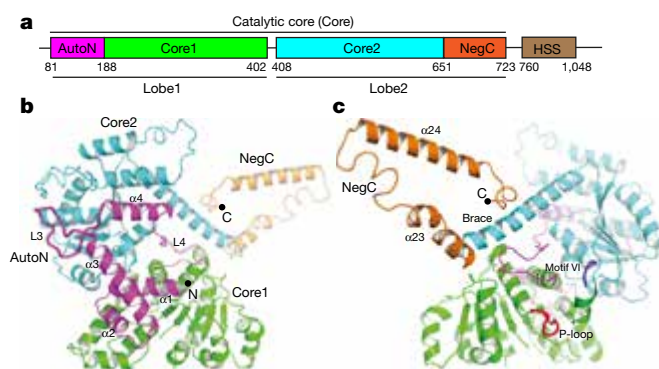


Figure 1 | Overall structure of MtISWI. **a**, Domain architecture of MtISWI. The catalytic core MtISWI (81–723; Core) used for crystallization consists of N-terminal AutoN (magenta), core1 (green), core2 (cyan), and C-terminal NegC (orange) domains. The C-terminal HSS domain was not used for crystallization and is coloured brown. **b**, **c**, Two different views of the overall structure of MtISWI. Domains are coloured as in Fig. 1a. The N and C termini of the protein are labelled. Motif I (P-loop) and motif VI are coloured red and blue, respectively, in **c**.

¹MOE Key Laboratory of Protein Science, Tsinghua University, Beijing, 100084, China. ²School of Life Science, Tsinghua University, Beijing, 100084, China.

*These authors contributed equally to this work.

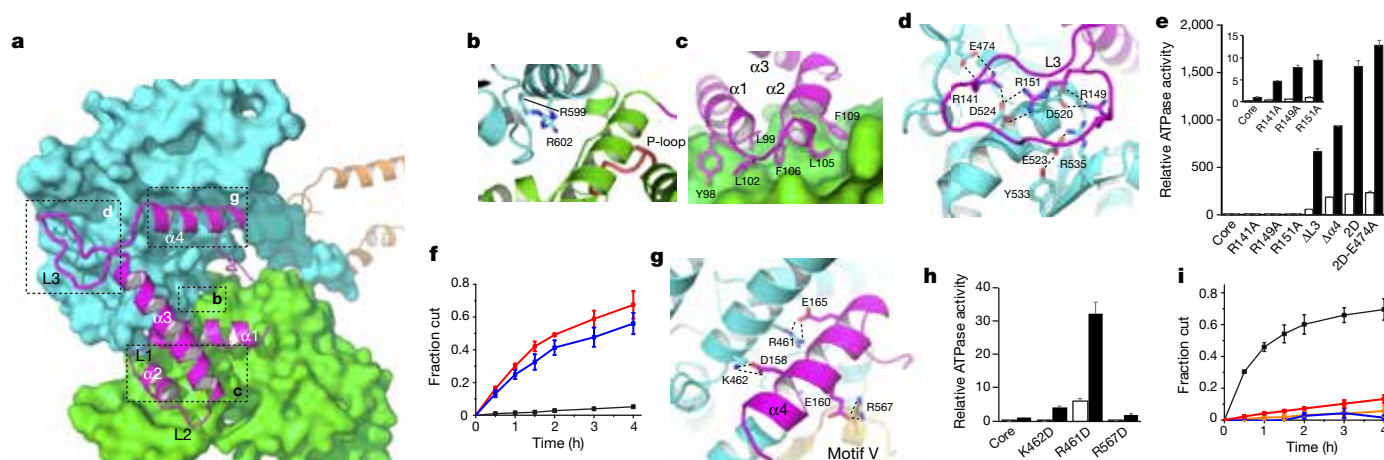


Figure 2 | Mechanism of MtiSWI inhibition by AutoN. **a**, Interaction between AutoN and the core domains of MtiSWI. The core1 and core2 domains are shown as surface presentations. The boxed regions are enlarged for further examination in **b–d** and **g**. **b**, Inactive conformation of ISWI. The arginine fingers of MtiSWI are labelled. **c**, Interaction between the N terminus of the AutoN domain and the core1 domain. **d**, Interaction between the H4-like L3 loop of the AutoN domain and the core2 domain. H-bond/salt-bridge interactions are shown as dotted lines. **e**, Relative ATPase activities of the catalytic core MtiSWI (81–723) with wild-type (WT) interface (Core) and various mutants in the presence (filled bars) and absence (open bars) of DNA. The specific activities of

the proteins were normalized to Core in the presence of DNA. Error bars, s.d. ($n = 3$). **f**, Chromatin remodelling activities of MtiSWI (81–723) with intact interface (Core; black), Δ L3 (red), and $\Delta\alpha 4$ (blue). Error bars, s.d. ($n = 3$). **g**, Interaction between $\alpha 4$ and the core2 domain of ISWI. Motif V is coloured gold. **h**, Relative ATPase activities of MtiSWI (81–723; Core) and three $\alpha 4$ -binding mutants. The assays were performed as in **e**. Error bars, s.d. ($n = 3$). **i**, Chromatin remodelling activities of MtiSWI (81–723) with intact interface (Core; black), R567D (red), R461D (brown), and K462D (blue). Owing to the low activity, the proteins were used at a high concentration (5 μ M). Error bars, s.d. ($n = 3$).

enzymes carrying the combined mutations displayed significantly higher ATPase and remodelling activities (Extended Data Fig. 4a–c). Replacement of the L3 loop with a flexible linker (L3) increased the ATPase and remodelling activities further (Fig. 2e, f and Extended Data Fig. 4d).

In addition to the H4-like L3 loop, the newly defined AutoN contains another inhibitory element, the $\alpha 4$ helix, which forms hydrogen bonds with Arg461, Lys462 and Arg567 of the core2 domain and mostly regulates ISWI in an H4 independent manner (Fig. 2g). Notably, similar to Δ L3, the $\alpha 4$ -deletion mutant ($\Delta\alpha 4$) showed ATPase and remodelling activities (Fig. 2e, f). The combined deletion of L3 and $\alpha 4$ (the 2D mutant) had an even more profound effect (Fig. 2e), suggesting AutoN utilizes multiple mechanisms to repress the activity of ISWI.

The $\alpha 4$ helix sequesters an important surface of the core2 domain. Mutations of the residues of the core2 domain that contacts $\alpha 4$ (R461D, K462D, and R567D) increased the ATPase activity (Fig. 2h), yet severely diminished the remodelling activity of ISWI (Fig. 2i and Extended Data Fig. 4e). This is not unexpected, as Arg567 of the core2 domain is an essential element of the canonical helicase motif V, which is important for the activities of Snf2 and Chd1 remodellers^{12–14}. These results suggest motif V and the nearby surface (including Arg461 and Lys462) that interact with $\alpha 4$ are not only important for ISWI inhibition, but are also involved in the remodelling process.

Interestingly, Arg567 is equivalent to Arg772 in motif V of ScChd1 (Extended Data Fig. 3f), which is masked by the double chromo-domain (dCD) and is essential for Chd1 inhibition¹². Thus, the mechanism of ISWI regulation by $\alpha 4$ is analogous to the Chd1 inhibition by dCD. In contrast, motif V of Snf2 is exposed to solvent and the enzyme is regulated differently^{13,15}, whereas the distally related protein SsoRad54 seems to lack any regulatory element¹⁶.

It was proposed that the H4 tail might bind the core1 domain and the C-terminal HSS domain^{8,17}. However, the binding mode of the H4-like L3 loop suggests H4 binds to the core2 domain. To visualize the recognition of H4, we determined the crystal structure of lobe2 of MtiSWI (406–754) bound with an H4 tail peptide (Fig. 3a and Extended Data Table 1). The structure shows clear electron density around Arg17 of the H4 peptide (Extended Data Fig. 5a). The L3- and H4-binding modes are not identical (Fig. 3b). Nevertheless, the H4-binding surface

is coincident with the L3 loop-binding site, suggesting H4-binding is incompatible with L3-mediated inhibition. This provides the structural basis of ISWI activation by the H4 tail.

The most striking feature of the H4-binding interface is the long side chain of Arg17 of H4, which is embedded in a negatively charged pocket composed of Glu474 and Asp524 (Fig. 3b). The binding of H4 is further stabilized through hydrogen bonds between the main chain amide groups of Lys16 and Arg17 of the peptide and the side chain of Asp524 of the core2 domain. The side chain of Lys16 packs against the binding interface and is proximal to Glu523. The conformation of the rest of the H4 peptide could not be defined in the crystals, consistent with the previous study showing that H4 tail stabilizes a particular conformation of the enzyme rather than acting as a mechanical element¹⁸. Supporting the structural model, the acetylated H4K16 peptide showed an approximate fourfold lower affinity (Extended Data Fig. 5d), suggesting that the charge-based interaction is important for H4 binding. Likewise, the R19A mutant caused a similar reduction in binding affinity (Extended Data Fig. 5e). The abilities of these H4 peptides to activate ISWI correlate with their binding strengths, with the acetylated peptide showing weaker stimulatory effect (Fig. 3d), consistent with the notion that fine tuning of the ISWI activity is required for the proper structure of the hyperacetylated X chromosome in the male fly^{2,22}.

The identification of the sites responsible for H4 recognition allows us to dissect their contributions. The E474A and D524A mutations severely weakened the binding of the H4 tail (Extended Data Fig. 6a), suggesting that Glu474 and Asp524 play important roles in H4 recognition. Notably, whereas the release of AutoN inhibition in the 2D mutant greatly enhanced the remodelling activity of the protein (Fig. 3e and Extended Data Fig. 6b), this activity was markedly attenuated by the H4-binding mutation E474A (the 2D-E474A mutant; Fig. 3e). The loss of remodelling activity was not due to a change in the overall structural integrity of the protein, as the 2D-E474A mutant retained an intact ATPase activity (Fig. 2e). Similarly, the 2D-D524A mutant also showed a marked loss of the remodelling activity (Fig. 3e). Likewise, in the context of the intact AutoN domain MtiSWI (81–723), the D524A

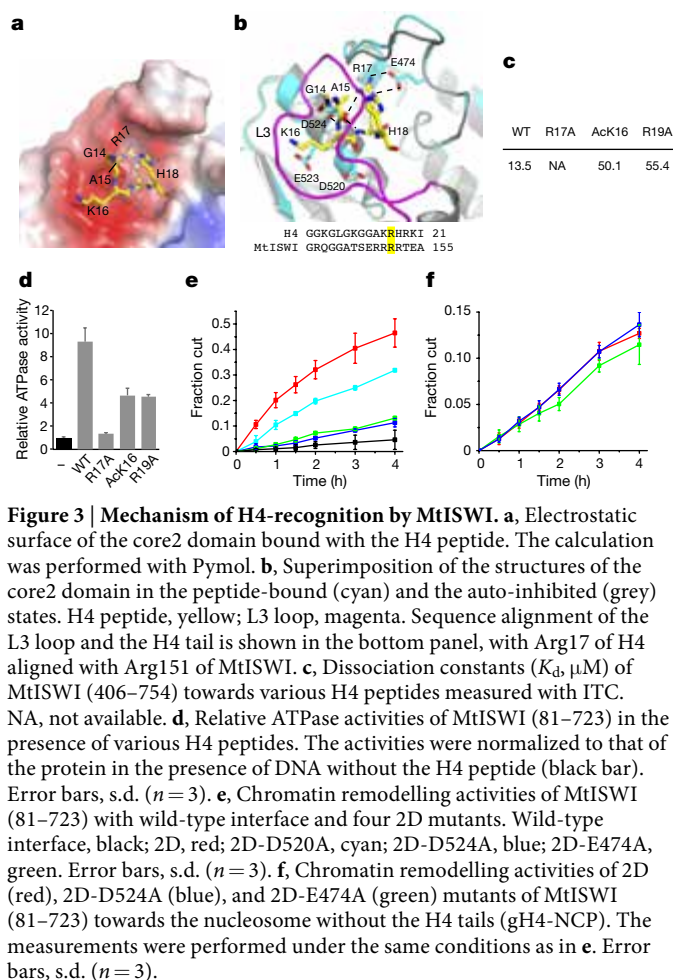


Figure 3 | Mechanism of H4-recognition by MtlISWI. **a**, Electrostatic surface of the core2 domain bound with the H4 peptide. The calculation was performed with Pymol. **b**, Superimposition of the structures of the core2 domain in the peptide-bound (cyan) and the auto-inhibited (grey) states. H4 peptide, yellow; L3 loop, magenta. Sequence alignment of the L3 loop and the H4 tail is shown in the bottom panel, with Arg17 of H4 aligned with Arg151 of MtlISWI. **c**, Dissociation constants (K_d , μ M) of MtlISWI (406–754) towards various H4 peptides measured with ITC. NA, not available. **d**, Relative ATPase activities of MtlISWI (81–723) in the presence of various H4 peptides. The activities were normalized to that of the protein in the presence of DNA without the H4 peptide (black bar). Error bars, s.d. ($n = 3$). **e**, Chromatin remodelling activities of MtlISWI (81–723) with wild-type interface and four 2D mutants. Wild-type interface, black; 2D, red; 2D-D520A, cyan; 2D-D524A, blue; 2D-E474A, green. Error bars, s.d. ($n = 3$). **f**, Chromatin remodelling activities of 2D (red), 2D-D524A (blue), and 2D-E474A (green) mutants of MtlISWI (81–723) towards the nucleosome without the H4 tails (gH4-NCP). The measurements were performed under the same conditions as in **e**. Error bars, s.d. ($n = 3$).

and E474A mutations released the auto-inhibition and stimulated the ATPase activity of the enzyme (Extended Data Fig. 6c), yet noticeably reduced the remodelling activity (Extended Data Fig. 6d). These results support our model in which the acidic surface patch of ISWI not only binds the L3 loop and inhibits ATP hydrolysis, but also binds the H4 tail and promotes chromatin remodelling.

To eliminate the possibility that residues Glu474 and Asp524 play a general catalytic role in the remodelling reaction, we removed the histone H4 tails (residues 1–20), assembled the mutant nucleosomes (gH4-NCPs), then measured the remodelling activity. The 2D mutant displayed a lower activity towards gH4-NCP relative to the intact NCP (Fig. 3f). Consistent with our model, disruption of the H4-binding

pocket did not further perturb the activity of the enzyme towards the gH4-NCP, with 2D-E474A and 2D-D524A showing comparable activity to the 2D mutant. Thus, Glu474 and Asp524 do not play a general catalytic role in the remodelling reaction. These findings indicate that the release of AutoN inhibition cannot bypass the requirement of the tight binding of the H4 tails for an efficient remodelling reaction, suggesting the H4 tails may function more than the antagonism of AutoN, probably orienting ISWI at SHL2 of the nucleosome⁷.

Intriguingly, the essential H4-binding residues are conserved in Snf2 subfamily remodellers (Extended Data Fig. 1), which also bind to SHL2 of the nucleosome^{23,24}, suggesting H4 may stabilize the binding of Snf2 at SHL2 through a conserved mechanism as in ISWI. However, different from ISWI, Snf2 is not locked into an inactive state (Extended Data Fig. 3a, b)¹³, and its activation is not strictly dependent on the H4 tail². These findings suggest that in addition to SHL2, Snf2 may bind to a different position of the nucleosome distal to the H4 tails, and thereby free of the regulation by the histone epitope.

ISWI is also regulated by NegC^{11,25}. NegC is preceded by the Brace helix (Fig. 4a), which functions in core1–core2 communication in many SF2 helicase proteins²⁶ and displays different numbers of helical turns in different remodellers (Extended Data Fig. 3f)^{12,13}. Supporting the importance of the Brace helix, mutation of the conserved Val638 residue of the Brace helix in the otherwise hyperactive 2D mutant (2D-V638D) resulted in a dramatic loss of the remodelling activity (Extended Data Fig. 2b). The Brace helix of ISWI protrudes from the C-terminal end of the core2 domain, and NegC extends further outwards and binds to the core2 domain of an adjacent molecule in the crystals, leading to formation of a domain-exchanged dimer (Fig. 4a). The NegC–core2 binding interface involves several highly conserved hydrophobic residues, including Ile661, Val668, Ile690, and Leu694 (Extended Data Fig. 1). This structure of NegC is different from the SnAc domain of Snf2 and C-terminal bridge of Chd1 (Extended Data Fig. 3h), which are involved in regulation of the remodellers^{12,13,27,28}.

ISWI is predominantly a monomer in the autoinhibited state^{6,29}, whereas dimer formation is markedly enhanced in the 2D mutant (Extended Data Fig. 2c), suggesting variable dimerization in the different mutants. The NegC–core2 interactions are not simply a crystal packing effect, as we observed the same interface in two different crystal forms (Extended Data Fig. 2d). The simplest interpretation of variable dimerization is that some of the intramolecular NegC–core2 interactions that the enzyme experiences in its reaction cycle can also form between monomers when protein concentration is very high. The significance of the intermolecular interactions in the absence of nucleosomes remains to be seen.

NegC inhibits the catalytic core of ISWI¹¹, which is essential for the nucleosome-centring activity of the enzyme^{11,25}. To test the function of the NegC–core2 interactions in linker DNA sensing, we made a minimal full-length MtlISWI (81–1048; mFL) that contains the HSS domain.

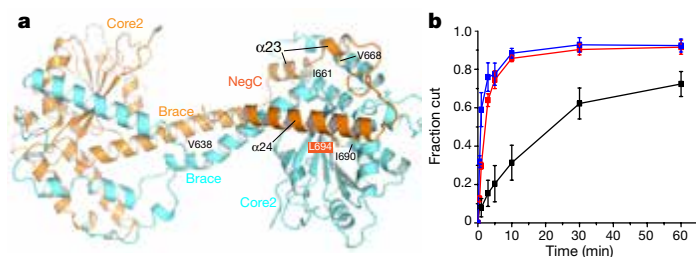
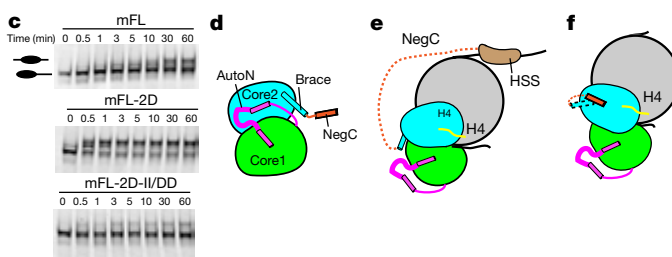


Figure 4 | NegC–core2 interactions and models for ISWI regulation.

a, Structure of ISWI dimer mediated by the NegC–core2 interactions in the crystals. One molecule is coloured cyan and the other orange. Val638 of the Brace helix and the residues involved in the NegC–core2 interactions are shown as stick presentations. **b**, Overall chromatin remodelling activities of mFL (black), mFL-2D (blue), and mFL-2D-II/DD (red). The assays were performed as in Fig. 2f. Error bars, s.d. ($n = 3$). **c**, Gels of the nucleosome-centring assays of mFL, mFL-2D and



mFL-2D-II/DD. The assays were performed similarly as described²⁵. Three sets of independent experiments were performed and the representative one is shown. **d**, Cartoon image of ISWI in the auto-inhibited state. The tethering HSS domain is omitted. **e**, Proposed model of ISWI bound to the nucleosome with long linker DNA. The dotted line illustrates NegC in the stressed condition. **f**, Proposed model of ISWI bound to the nucleosome with short linker DNA, in which the Brace helix is partly disrupted and NegC binds to core2 intramolecularly. The HSS domain is omitted.

mFL formed a low fraction of dimer in solution (Extended Data Fig. 2c), and the presence of the HSS domain conferred robust ATPase, overall chromatin remodelling and nucleosome-centring activities (Fig. 4b, c and Extended Data Fig. 2e, f). Similar to what we showed above, the release of AutoN inhibition by deletions of L3 and $\alpha 4$ of mFL (mFL-2D) greatly enhanced the activities of the enzyme, suggesting that AutoN plays a major role in ISWI inhibition, but is not required for linker DNA sensing. In sharp contrast, disruption of the hydrophobic NegC–core2 interactions by I661D I690D mutations (mFL-2D-II/DD) resulted in a large loss of the centring activity (Fig. 4c), with little change in the ATPase and overall nucleosome remodelling activities (Fig. 4b and Extended Data Fig. 2e, f). These results suggest that the NegC–core2 interactions are essential for the regulation of linker DNA sensing in the context of the full-length protein. Considering the plasticity of the Brace helix shown in different remodellers, we propose that NegC may fold back and bind to the core2 domain intramolecularly when ISWI engages with the nucleosome, leading to partial melting of the preceding Brace helix and inhibition of the remodelling activity.

Taken together, our findings reveal how ISWI is autoinhibited and provide plausible mechanisms for its regulation by nucleosomal epitopes. In the absence of the nucleosome, two essential elements (L3 and $\alpha 4$) of AutoN regulate ISWI by collectively holding the catalytic core in an inactive conformation (Fig. 4d). Upon engagement with the nucleosome with long linker DNA, the HSS domain of ISWI stably binds to the extranucleosomal linker DNA and is located distal to the core2 domain, which binds the H4 tail and the nucleosomal DNA at SHL2 (Fig. 4e). In this conformation, the structure of NegC is disturbed, and thereby allows an efficient remodelling reaction. With short linker DNA, the HSS domain does not, or only weakly, binds to the extranucleosomal DNA, freeing NegC to interact with the core2 domain and induce an inactive conformation (Fig. 4f). Thus, NegC works as a brake to stop the remodeller when encountering a nucleosome with short linker DNA, conferring the enzyme with nucleosome-centring activity. It has recently been proposed that the HSS domain binds the nucleosome core in a translocation competent state²⁵. How the HSS domain interacts with the nucleosome core and regulates NegC in this state is currently unknown; more study is needed.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 April; accepted 27 October 2016.

Published online 5 December 2016.

- Manning, B. J. & Peterson, C. L. Releasing the brakes on a chromatin-remodeling enzyme. *Nature Struct. Mol. Biol.* **20**, 5–7 (2013).
- Clapier, C. R. & Cairns, B. R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).
- Varga-Weisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388**, 598–602 (1997).
- Tsukiyama, T., Daniel, C., Tamkun, J. & Wu, C. ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. *Cell* **83**, 1021–1026 (1995).
- Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145–155 (1997).
- Mueller-Planitz, F., Klinker, H., Ludwigsen, J. & Becker, P. B. The ATPase domain of ISWI is an autonomous nucleosome remodeling machine. *Nature Struct. Mol. Biol.* **20**, 82–89 (2013).
- Dang, W., Kagalwala, M. N. & Bartholomew, B. Regulation of ISW2 by concerted action of histone H4 tail and extranucleosomal DNA. *Mol. Cell. Biol.* **26**, 7388–7396 (2006).
- Grüne, T. *et al.* Crystal structure and functional analysis of a nucleosome recognition module of the remodeling factor ISWI. *Mol. Cell* **12**, 449–460 (2003).
- Hota, S. K. *et al.* Nucleosome mobilization by ISW2 requires the concerted action of the ATPase and SLIDE domains. *Nature Struct. Mol. Biol.* **20**, 222–229 (2013).

- Ludwigsen, J., Klinker, H. & Mueller-Planitz, F. No need for a power stroke in ISWI-mediated nucleosome sliding. *EMBO Rep.* **14**, 1092–1097 (2013).
- Clapier, C. R. & Cairns, B. R. Regulation of ISWI involves inhibitory modules antagonized by nucleosomal epitopes. *Nature* **492**, 280–284 (2012).
- Hauk, G., McKnight, J. N., Nodelman, I. M. & Bowman, G. D. The chromodomains of the Chd1 chromatin remodeler regulate DNA access to the ATPase motor. *Mol. Cell* **39**, 711–723 (2010).
- Xia, X., Liu, X., Li, T., Fang, X. & Chen, Z. Structure of chromatin remodeler Swi2/Snf2 in the resting state. *Nature Struct. Mol. Biol.* **23**, 722–729 (2016).
- Smith, C. L. & Peterson, C. L. A conserved Swi2/Snf2 ATPase motif couples ATP hydrolysis to chromatin remodeling. *Mol. Cell. Biol.* **25**, 5880–5892 (2005).
- Clapier, C. R. *et al.* Regulation of DNA translocation efficiency within the chromatin remodeler RSC/Sth1 potentiates nucleosome sliding and ejection. *Mol. Cell* **62**, 453–461 (2016).
- Dürr, H., Körner, C., Müller, M., Hickmann, V. & Hopfner, K. P. X-ray structures of the *Sulfolobus solfataricus* SWI2/SNF2 ATPase core and its complex with DNA. *Cell* **121**, 363–373 (2005).
- Dang, W. & Bartholomew, B. Domain architecture of the catalytic subunit in the ISW2-nucleosome complex. *Mol. Cell. Biol.* **27**, 8306–8317 (2007).
- Racki, L. R. *et al.* The histone H4 tail regulates the conformation of the ATP-binding pocket in the SNF2h chromatin remodeling enzyme. *J. Mol. Biol.* **426**, 2034–2044 (2014).
- Corona, D. F., Clapier, C. R., Becker, P. B. & Tamkun, J. W. Modulation of ISWI function by site-specific histone acetylation. *EMBO Rep.* **3**, 242–247 (2002).
- Clapier, C. R., Nightingale, K. P. & Becker, P. B. A critical epitope for substrate recognition by the nucleosome remodeling ATPase ISWI. *Nucleic Acids Res.* **30**, 649–655 (2002).
- Ferreira, H., Flaus, A. & Owen-Hughes, T. Histone modifications influence the action of Snf2 family remodelling enzymes by different mechanisms. *J. Mol. Biol.* **374**, 563–579 (2007).
- Deuring, R. *et al.* The ISWI chromatin-remodeling protein is required for gene expression and the maintenance of higher order chromatin structure *in vivo*. *Mol. Cell* **5**, 355–365 (2000).
- Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodeling through directional DNA translocation from an internal nucleosomal site. *Nature Struct. Mol. Biol.* **12**, 747–755 (2005).
- Zofall, M., Persinger, J., Kassabov, S. R. & Bartholomew, B. Chromatin remodeling by ISW2 and SWI/SNF requires DNA translocation inside the nucleosome. *Nature Struct. Mol. Biol.* **13**, 339–346 (2006).
- Leonard, J. D. & Narlikar, G. J. A nucleotide-driven switch regulates flanking DNA length sensing by a dimeric chromatin remodeler. *Mol. Cell* **57**, 850–859 (2015).
- Fairman-Williams, M. E., Guenther, U. P. & Jankowsky, E. SF1 and SF2 helicases: family matters. *Curr. Opin. Struct. Biol.* **20**, 313–324 (2010).
- Sen, P., Ghosh, S., Pugh, B. F. & Bartholomew, B. A new, highly conserved domain in Swi2/Snf2 is required for SWI/SNF remodeling. *Nucleic Acids Res.* **39**, 9155–9166 (2011).
- Sen, P. *et al.* The SnAC domain of SWI/SNF is a histone anchor required for remodeling. *Mol. Cell. Biol.* **33**, 360–370 (2013).
- Racki, L. R. *et al.* The chromatin remodeler ACF acts as a dimeric motor to space nucleosomes. *Nature* **462**, 1016–1021 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Fan at the centre of structure biology (Tsinghua University) and the staff at beamline BL17U of Shanghai Synchrotron Radiation Facility for help with diffraction data collection, and the Tsinghua University Branch of the China National Center for Protein Sciences Beijing for providing facility support. This work was supported by the Chinese Key Research Plan-Protein Sciences (2014CB910100), the National Natural Science Foundation of China (31570731, 31270762), and the 'Junior One Thousand Talents' program to Z.C.

Author Contributions L.Y. and L.W. prepared the proteins and performed the biochemical analyses with the help from X.X. and Y.T.; L.Y. crystallized the proteins; Z.C. wrote the manuscript with help from all authors; Z.C. directed and supervised all the research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.C. (Zhucheng_chen@tsinghua.edu.cn).

Reviewer Information Nature thanks B. Bartholomew and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification. We cloned the gene of MtISWI from *M. thermophila* genomic DNA, then inserted it between the NdeI and NotI restriction sites in a modified pGEX-4T-2 (GST-tag) vector, in which the thrombin recognition site was replaced with tobacco etch virus (TEV) protease recognition site. After expression test and several trials, we selected the catalytic core of ISWI (81–723) as the construct for crystallization and most of the biochemical analyses. MtISWI (406–754) was used in co-crystallization with the H4 peptide. To generate the Δ L3, Δ α4, and the 2D mutants, we replaced the residues 137–152 of the L3 loop, the residues 154–171 of α4, and the residues 137–171 of AutoN with a flexible linker sequence of GGASGS, SGASGG, and GGSGSGSGSGSGGS, respectively. To test the nucleosome-centring activity of ISWI, we made a minimal full-length MtISWI (81–1048; mFL). To facilitate purification of the mFL-2D-II/DD mutant, a Strep-tag was added to the C terminus of the protein. Various point mutants were generated by Quickchange mutagenesis. All constructs were confirmed by DNA sequencing.

ISWI (81–723) was overexpressed in the Chaperon Competent Cell pG-Tf2 BL21 (DE3) strain of *Escherichia coli* and cells were grown in LB media to an absorbance at 600 nm ($A_{600\text{nm}}$) = 0.8 at 37 °C and induced with 10 ng/ml tetracycline and 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) at 18 °C overnight. Cells were spun down at speed 4,000 r.p.m. (Beckman, Rotor JA4.2) for 15 min, and lysed by High Pressure Homogenizer Machine (ATS) at 4 °C, in 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 5 mM EDTA, 2 mM dithiothreitol (DTT), and 1 mM PMSE. The cell lysate was cleared by centrifugation at speed 19,000 r.p.m. (Beckman, Rotor JA20) for 1 h, the supernatant was separated and then loaded onto a gravity column with GST beads. After elution by 25 mM reduced glutathione (GSH), the GST-tagged fusion protein was cleaved by incubating with ~0.02 mg/ml TEV protease overnight at 4 °C. The proteins were further purified by ion-exchange (source Q, GE Healthcare) and gel-filtration (Superdex200, GE Healthcare) chromatography. ISWI (81–723) was concentrated to 15 mg/ml before crystallization in 10 mM HEPES (pH 7.5), 150 mM NaCl, and 10 mM dithiothreitol. The other proteins of MtISWI (81–723) with point mutations (R141A, R149A, R151A, 2RA, 3RA, R461D, K462D, R567D, E474A, D520A, and D524A) were purified similarly, and concentrated to 5–10 mg/ml for assays.

The ISWI mutants (2D, 2D-V638D, mFL, mFL-2D, and mFL-2D-II/DD) are not very stable, so were purified slightly differently by adding 1 mM MgCl₂, 0.5 mM ATP, 10% glycerol to the lysis buffer. The GST tag was removed after ion-exchange purification by treatment with TEV protease. mFL-2D-II/DD was purified with an additional step of Strep-tag affinity column before being subjected to gel-filtration chromatography. The purified proteins were stored in 10 mM HEPES (pH 7.5), 300 mM NaCl, 10% glycerol, and 10 mM dithiothreitol.

ISWI (406–754) was purified by ion-exchange (Source S, GE Healthcare) with HEPES buffer (pH 7.0) and gel-filtration (Superdex75, GE Healthcare) chromatography. It was concentrated to 20 mg/ml before crystallization in 10 mM HEPES (pH 7.5), 150 mM NaCl, and 10 mM DTT.

The H4-tail peptide used for co-crystallization contains residues Ser1 to Ile 21. The H4-tail peptides used for ATPase assay and ITC measurements encompass residues Lys12 to Lys20. They were purchased from Scilight Biotechnology (98% purity).

Crystallization and data collection. Crystals of ISWI (81–723) were grown at 4 °C by hanging drop vapour diffusion above a reservoir solution of 200 mM NaAc (pH 4.8), 16% PEG3350, 10 mM dithiothreitol, with protein to reservoir volume ratio 1:1. Specific drops yielded crystals with dimensions of 100 μm × 200 μm × 200 μm. Crystals were harvested in cryo-protectant containing 10–25% w/v glycerol and then flash-frozen in liquid nitrogen. For co-crystallization with the H4 tail, approximately 500 μM ISWI (406–754) was incubated with 2 mM H4-tail peptide. Crystals of the ISWI (406–754)–H4 complex were grown from 0.03 M citric acid/0.07 M BIS-TRIS propane (pH 7.6), 20% w/v polyethylene glycol 400, 3% w/v trimethylamine N-oxide dehydrate, 10 mM dithiothreitol, with equal volumes of protein and reservoir buffer. Crystals were harvested in reservoir buffer with maximally 35% w/v polyethylene glycol 400 as cryo-protectant and then flash-frozen in liquid nitrogen.

Diffraction data from crystals of ISWI (81–723) and ISWI (406–754)–H4 complex were collected at –170 °C at the beamline BL17U of Shanghai Synchrotron Radiation Facility.

Data processing and structure solution. All the data were processed with the HKL2000. The structure of ISWI (81–723) was solved by molecular replacement using the core1 and core2 domains of Chd1 (Protein Data Bank accession number 3MWY) as the initial searching models. The rest of the model was built manually using Coot. Refinement was performed with Phenix³⁰. The final structure

was refined to 2.4 Å, with $R_{\text{work}}/R_{\text{free}} = 0.195/0.226$, Ramachandran outlier 0.0%, allowed 2.7%, and favoured 97.3%.

The structure of ISWI (406–754)–H4 complex was solved by molecular replacement using the core2 domain of MtISWI as the initial searching model. One asymmetric unit of the crystals contains eight copies of a dimer. After the structure of the first dimer was solved, it was then used as the search model to find the rest of the molecules. The final structure was refined to 3.0 Å, with $R_{\text{work}}/R_{\text{free}} = 0.218/0.276$, Ramachandran outlier 0.16%, allowed 3.6%, and favoured 96.2%.

Multi-angle light scattering (MALS). Protein (100 μl) at 1 mg/ml was injected into a Superdex200 column (GE Healthcare) equilibrated with the running buffer containing 10 mM HEPES (pH 7.5), 300 mM NaCl, 2 mM dithiothreitol. To stabilize mFL, mFL-2D, and mFL-2D-II/DD, 10% glycerol was added in the running buffer. The chromatography system was coupled to an 18-angle light scattering detector (Wyatt Technology) for data collection. Data were collected every 0.5 s at a flow rate of 0.5 ml/min. Bovine serum albumin (BSA, 68 kDa, Sigma) was used as a standard to normalize the system. Data analysis used program ASTRA 6.1.

ITC measurements. Protein ISWI (406–754) and the H4-tail peptides were kept in the same buffer containing 10 mM HEPES (pH 7.5), 150 mM NaCl. The concentrations of the H4-tail peptides (1,130 μM) in the syringe were about 11-fold higher than the concentration of ISWI in the cell (100 μM). As a control, the H4-tail peptides were injected into the reaction buffer without the protein, and the data were used to account for the heat of mixing/dilution. All the experiments were performed at 25 °C. The data were fitted in the Origin 7.0 software package of MicroCal-ITC implementation, yielding the dissociation constants (K_D) and reaction stoichiometry (n).

ATPase assays. Measurement of ATP hydrolysis was based on a spectrophotometric shift in the maximum absorbance of the substrate from 330 nm to 360 nm, resulting from the enzymatic conversion of 2-amino-6-mercapto-7-methylpurine riboside (MESG) by purine nucleoside phosphorylase (PNP) in the presence of Pi (EnzChek Phosphate Assay Kit). The measurements were performed in a Microplate Reader (VARIOSKAN FLASH, Thermo Scientific), and the ATPase activities were calculated at the early time points when the yield of product increased linearly.

ATPase assays were performed in 2 mM ATP and 2 μM dsDNA (25 bp), 50 mM Tris-HCl (pH 7.5), 1 mM MgCl₂, 0.2 mM sodium azide, 0.2 mM dithiothreitol, 15 mM sodium chloride. Owing to the auto-inhibited nature and the low activities, 1 μM of various MtISWI (81–723) proteins, including the enzymes with wild-type interface and with various point mutations, were used. Owing to the very high ATPase activities caused by the release of AutoN inhibition, 0.1 μM ISWI mutants (2D, 2D-E474A and 2D-V638D) were used. The specific ATPase activities of all the proteins were normalized to ISWI (81–723) with DNA. To stabilize the mutant proteins (mFL, mFL-2D, and mFL-2D-II/DD), the ATPase assays were performed in the presence of 50 mM sodium chloride and 10% glycerol, and normalized to the activity of MtISWI (81–723) under the same conditions. To measure the ATPase activities of the proteins with the HSS domain (mFL, mFL-2D, and mFL-2D-II/DD), 2 μM longer DNA (146 bp) and 0.1 μM ISWI were used. To measure the ATPase activity in the presence of H4 peptide, 1 μM MtISWI, 2 μM dsDNA, and 16 μM various H4 peptides were used.

Nucleosome remodelling assays. Mononucleosome restriction enzyme accessibility assays were performed as described³¹. Cy5-labelled mononucleosome (3 nM) and 0.2 μM of various ISWI proteins (5 μM proteins were used in Fig. 2i and Extended Data Figs 4b and 6d) were incubated at 37 °C with 3 mM ATP and 100 U of HhaI in the remodelling buffer (20 mM Tris, pH 7.5, 50 mM KCl, 3 mM MgCl₂, 10% glycerol, and 0.1 mg/ml bovine serum albumin). Fractions were taken at various time points and quenched with 2 × Stop buffer (20 mM Tris, pH 8.0, 0.6% sodium dodecyl sulphate (SDS), 40 mM EDTA, and 0.1 mg/ml proteinase K). The reaction mixtures were incubated at 55 °C for 20 min to deproteinate the samples. Fractions were running on 8% native TBE polyacrylamide gels in 0.25 × TBE for 120 min at 80 V on ice. Gels were imaged using a Typhoon 9410 variable mode imager (GE Healthcare). Band intensities were quantified in Quantity One software.

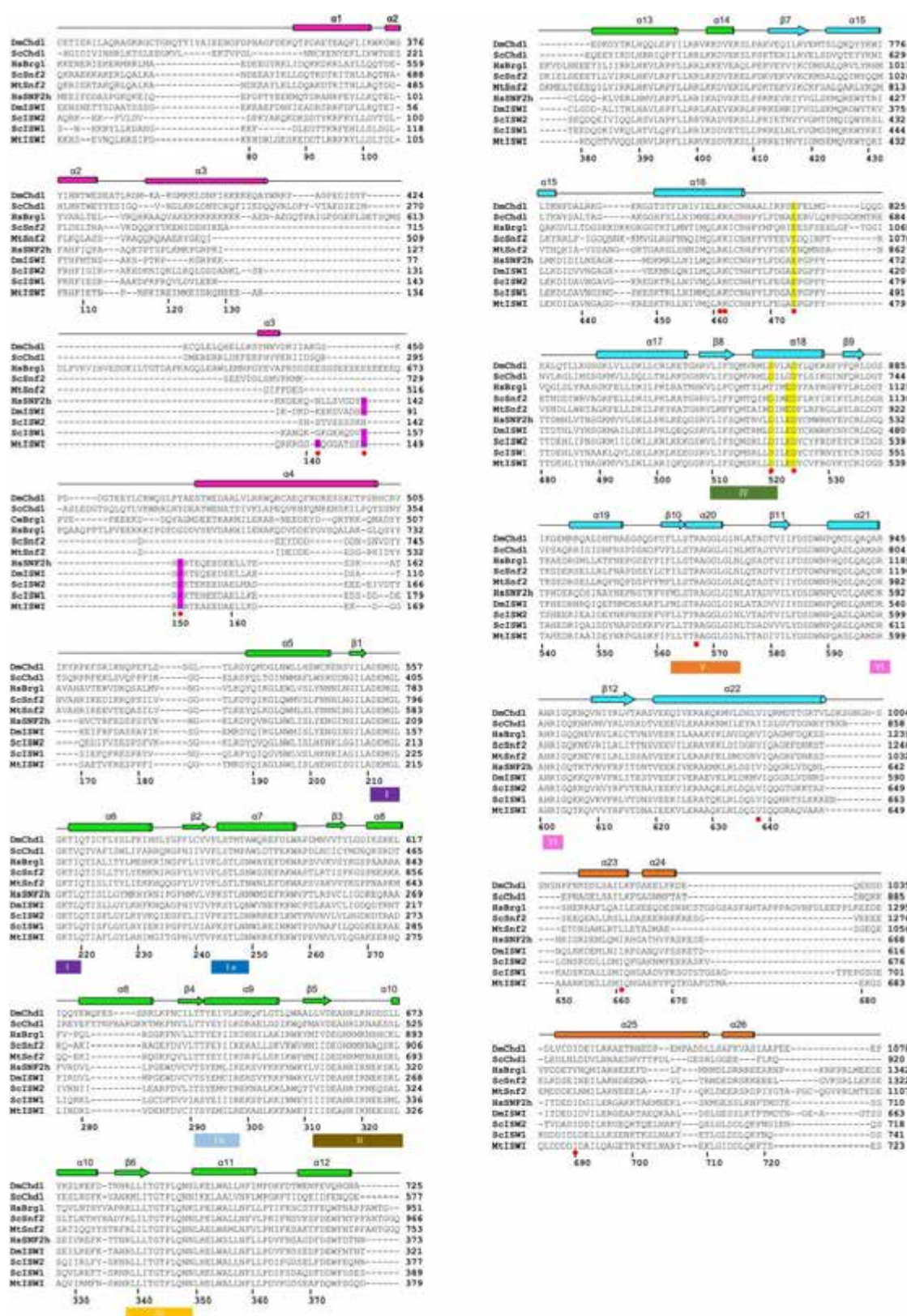
Nucleosome centring assays. Nucleosome centring assays were performed basically as described²⁵ with 40 nM NCP and 100 nM MtISWI at 37 °C. Reactions were performed in the buffer (70 mM KCl, 15 mM HEPES, 5 mM MgCl₂, 0.02% NP40, 10% glycerol, 1 mM dithiothreitol, 0.5 mM EDTA, pH 7.5) with 2 mM ATP, and stopped with 150 ng sperm DNA at the indicated time points. The products were resolved with 8% native acrylamide gels, 0.25 × TBE at 4 °C for 120 min at 150 V. The positions of fluorescently labelled DNA were detected using a Typhoon 9410 imager (GE Healthcare).

GST–H4 pull-down assays. GST-tagged H4 tail (SGRGKGGKGLGKGG AKRHRKI, 20 μM) was pre-incubated with GST resins, then mixed with 20 μM proteins on a rotating platform in the binding buffer (20 mM HEPES, 50 mM NaCl, and 2 mM DTT, pH 7.5) at 4 °C for 30 min. The GST resins were washed three times with the binding buffer and eluted with 25 mM glutathione (pH 8.0). Samples were

boiled in SDS loading buffer and analysed with 12% SDS–PAGE and Coomassie blue staining.

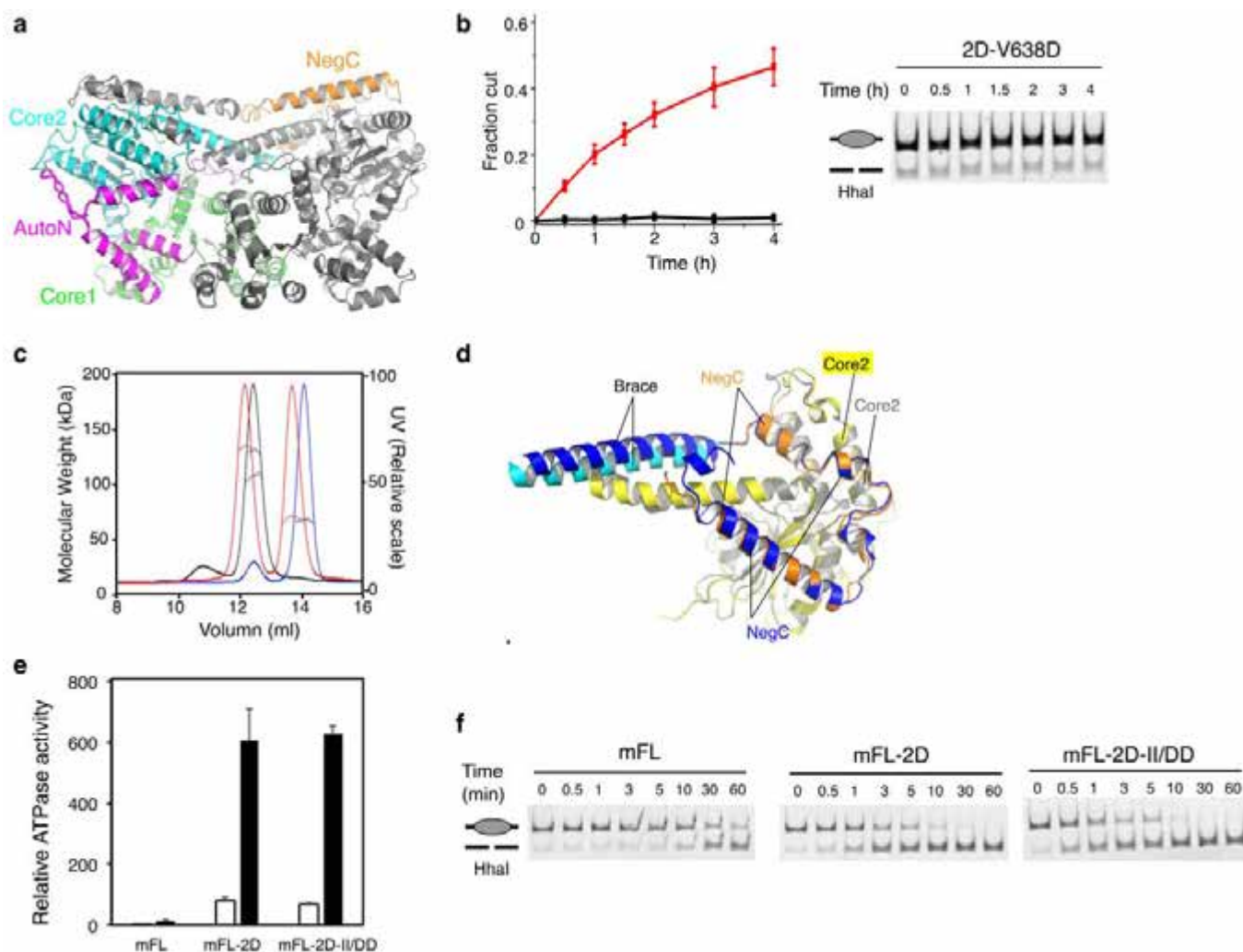
Data availability. Coordinates and structure factors have been deposited in the Protein Data Bank under accession numbers 5JXR (MtISWI) and 5JXT (MtISWI in complex with H4 peptide). All other data are available from the corresponding author upon reasonable request.

30. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
31. Yang, X., Zaurin, R., Beato, M. & Peterson, C. L. Swi3p controls SWI/SNF assembly and ATP-dependent H2A–H2B displacement. *Nature Struct. Mol. Biol.* **14**, 540–547 (2007).



Extended Data Figure 1 | Multiple sequence alignments of Chd1, Snf2, and ISWI subfamily of chromatin remodellers. The sequence alignments were done with Clustal Omega. Secondary structural assignments on the top are based on the structure determined in this study and colour coded as in Fig. 1a. The residue numbering at the bottom is based on the

sequence of MtiSWI. The helix motifs are assigned as reported¹³. The basic residues involved in AutoN inhibition are highlighted in magenta, and the acidic residues implicated in H4-binding are highlighted in yellow. The residues analysed in this study are indicated with red circles.

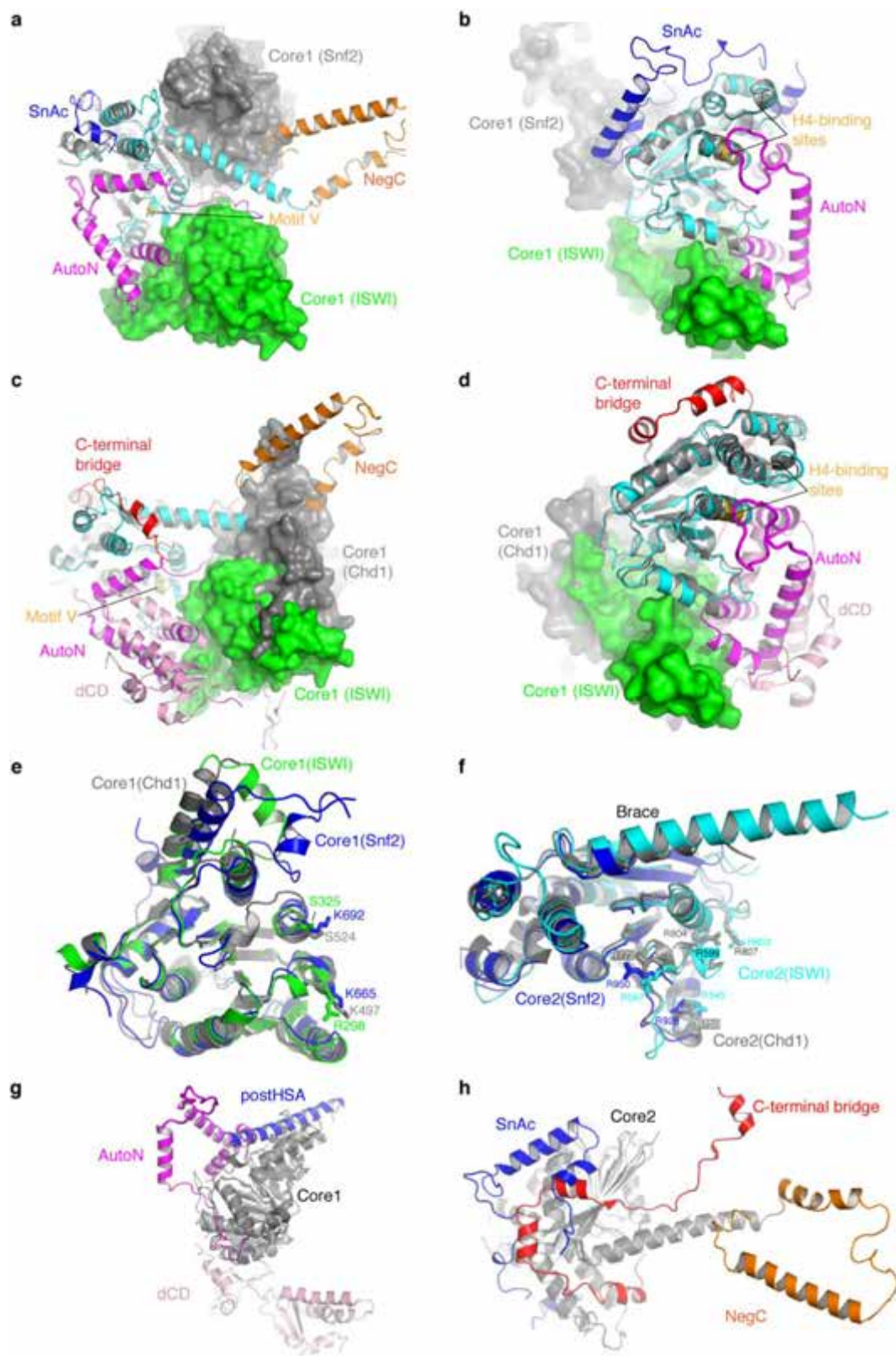


Extended Data Figure 2 | Analyses of ISWI regulation by NegC.

a, Structure of the ISWI dimer. One molecule is coloured as in Fig. 1, and the other molecule is coloured grey. **b**, Chromatin remodelling of the 2D-V638D mutant (black). The activity of the parental 2D mutant was reproduced from Fig. 3e (red). Error bars, s.d. ($n = 3$). One representative gel was shown in the right panel. **c**, MALS of the core MtISWI (81–723; blue), 2D mutant (red), and mFL (81–1048; black). Core and mFL MtISWI are predominantly in a monomeric state, with a small fraction of dimer (~6%). The calculated molecular masses of the major peaks of core and mFL MtISWI are ~66 kDa and ~106 kDa, respectively, corresponding to a monomer, whereas the small peaks correspond to the dimer fractions. The 2D mutant shows two peaks with molecular masses of ~68 kDa

and ~133 kDa, corresponding to a monomer and a dimer, respectively.

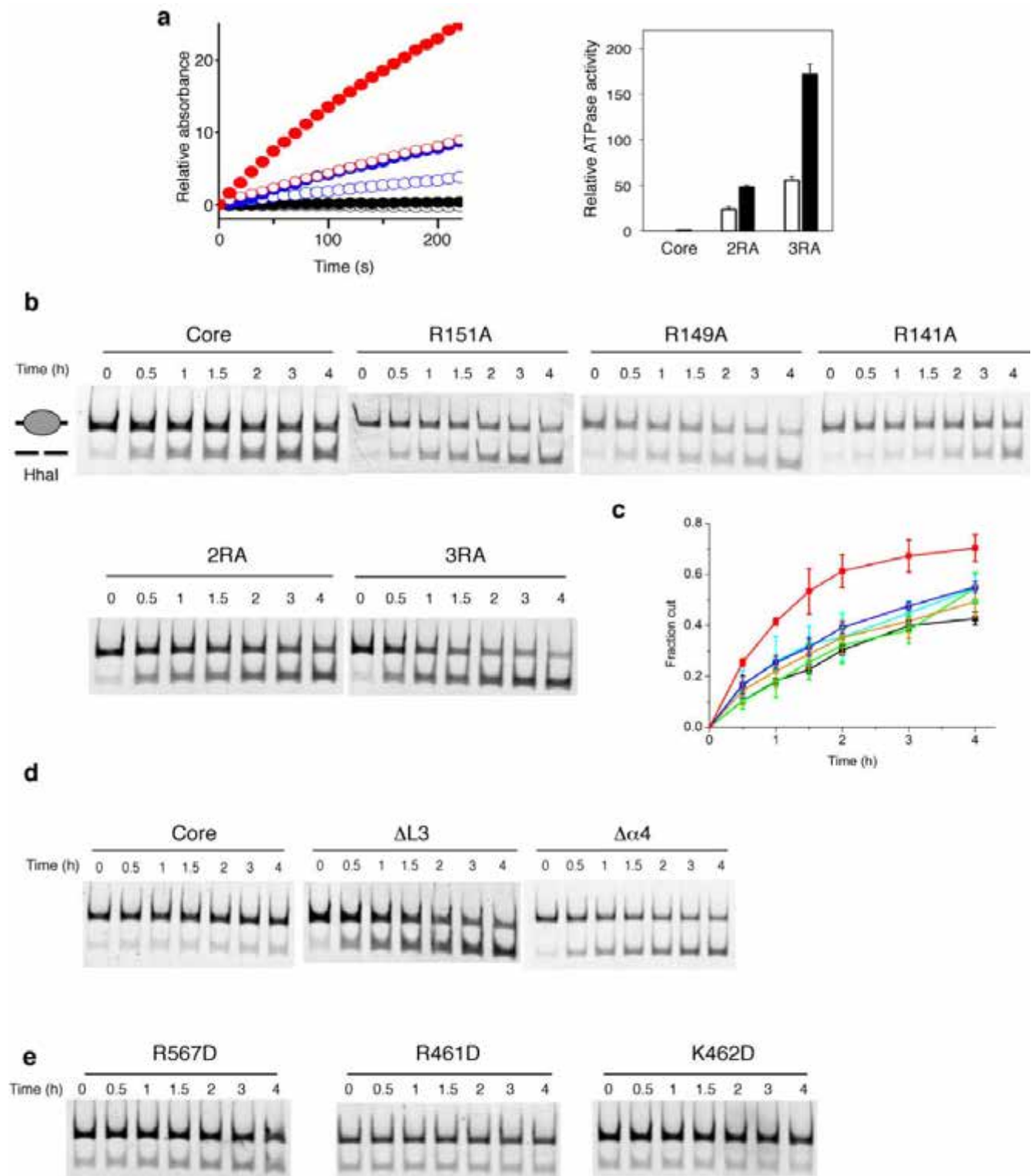
d, Superimposition of the core2 domains of the two crystal structures examined in this study. For clarity, only the dimerization interface is shown. One dimer is coloured as in **a**; the other dimer is coloured yellow and blue. The core2 and NegC domains interact similarly in the two different crystal forms. The Brace helix shows some domain movement relative to NegC. **e**, ATPase activities of mFL, mFL-2D, and mFL-2D-II/DD in the absence (open bars) and the presence (filled bars) of DNA. Error bars, s.d. ($n = 3$). **f**, Representative gels of the overall chromatin remodelling assays of mFL, mFL-2D, and mFL-2D-II/DD. Quantifications of the cut fractions are shown in Fig. 4b.



Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Structural comparisons among Snf2, Chd1, and ISWI. **a, b,** Comparisons of the overall structures of MtISWI and MtSnf2 (Protein Data Bank accession number 5HZR)¹³. The structures of the core2 domains are aligned. The core1 domains are shown as surface presentations, which orient differently in these two proteins. ISWI is coloured as in Fig. 1. The core1 and SnAc domains of MtSnf2 are coloured grey and blue, respectively. Motif V (R567 of MtISWI and R950 of MtSnf2) and the acidic patch of the core2 domain implicated in H4-binding are coloured gold. **c, d,** Comparisons of the overall structures of MtISWI and ScChd1 (Protein Data Bank accession number 3MWY)¹². The structures of the core2 domains are aligned. The core1 domains are shown as surface presentations. The N-terminal dCD and the C-terminal bridge of ScChd1 are coloured pink and red, respectively. The NegC domain of ISWI extends outwards, whereas the C-terminal bridge of Chd1 binds to the core1 domain intramolecularly. **e,** Structural alignment of the core1 domains

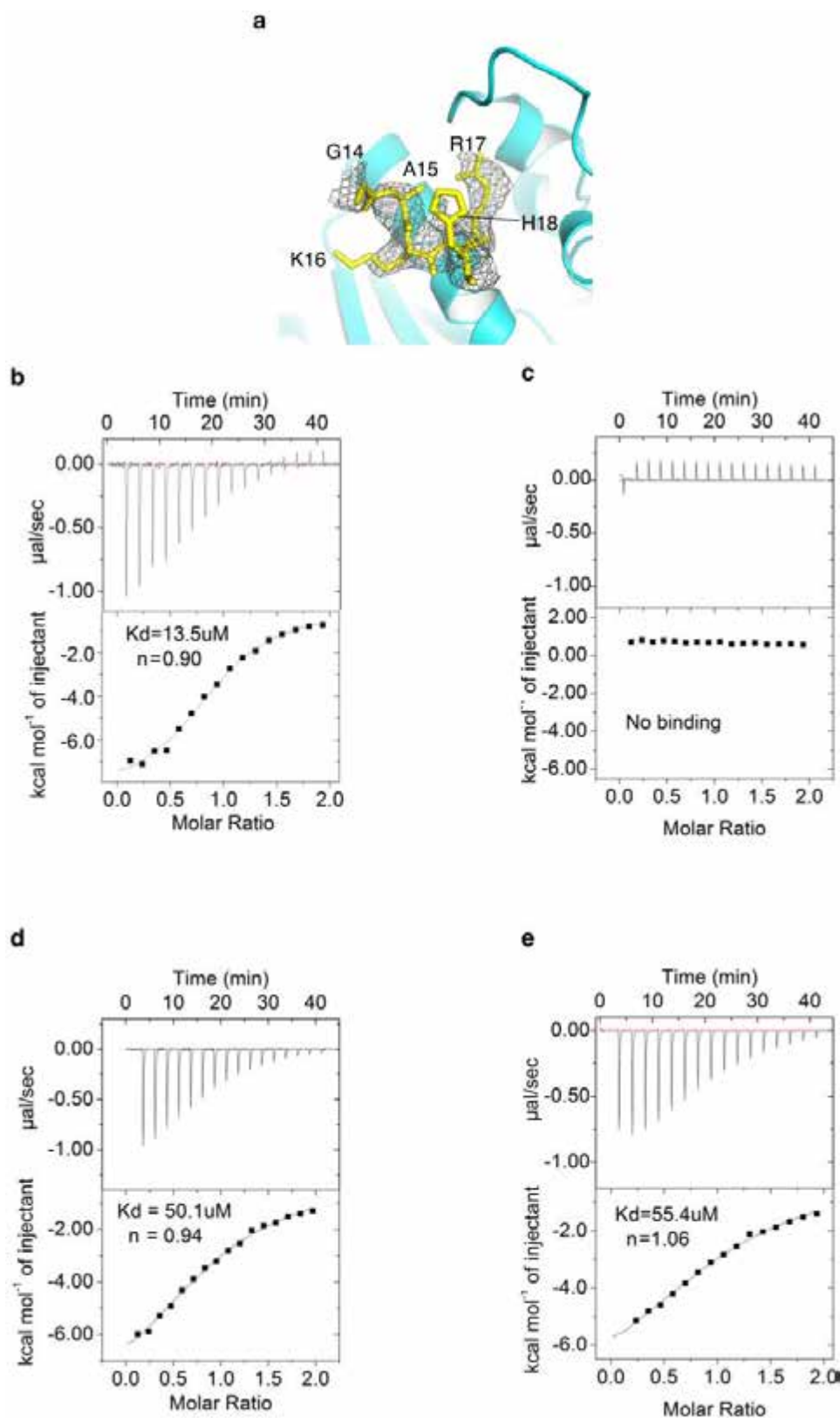
of MtISWI (green), MtSnf2 (blue), and ScChd1 (grey). **f,** Structural alignment of the core2 domains of MtISWI (cyan), MtSnf2 (blue), and ScChd1 (grey). The DNA-binding elements identified in MtSnf2 (K662 and R950) and ScChd1 (R750 and R772) are conserved among these remodelers, whereas K692 of MtSnf2 is unique to the Snf2-subfamily proteins. The arginine-fingers of MtISWI (R599 and R602), MtSnf2 (R982 and R985), and ScChd1 (R804 and R807) are conserved (Extended Data Fig. 1). The Brace helices of the remodelers show different lengths. **g,** Comparisons of lobe1 of MtISWI, MtSnf2 and ScChd1. The structures of the core1 domains are aligned. The N-terminal auxiliary domains of MtISWI (AutoN), MtSnf2 (postHSA), and ScChd1 (dCD) interact with the core1 domain differently. **h,** Comparisons of lobe2 of MtISWI, MtSnf2, and ScChd1. The structures of the core2 domains are aligned. The C-terminal auxiliary domains of MtISWI (NegC), MtSnf2 (SnAC), and ScChd1 (bridge) interact with the core2 domain differently.



Extended Data Figure 4 | Analyses of ISWI regulation by AutoN.

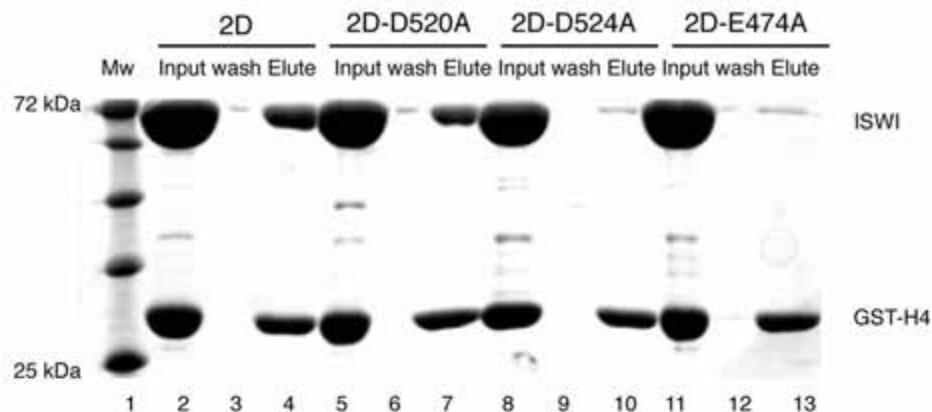
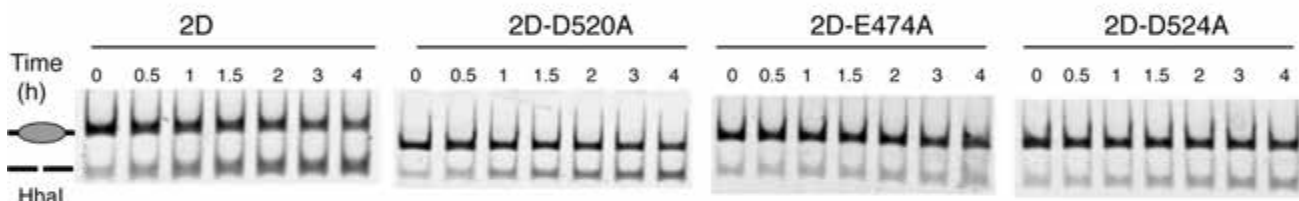
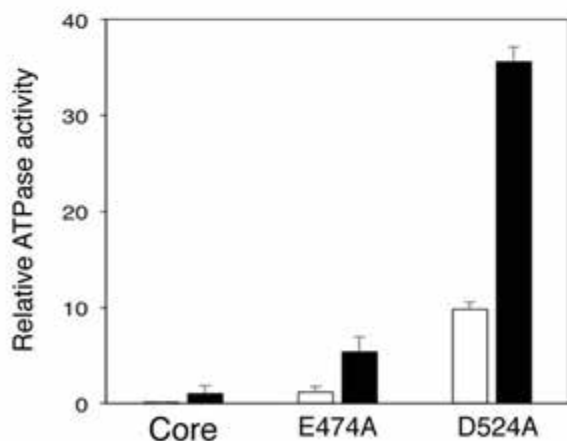
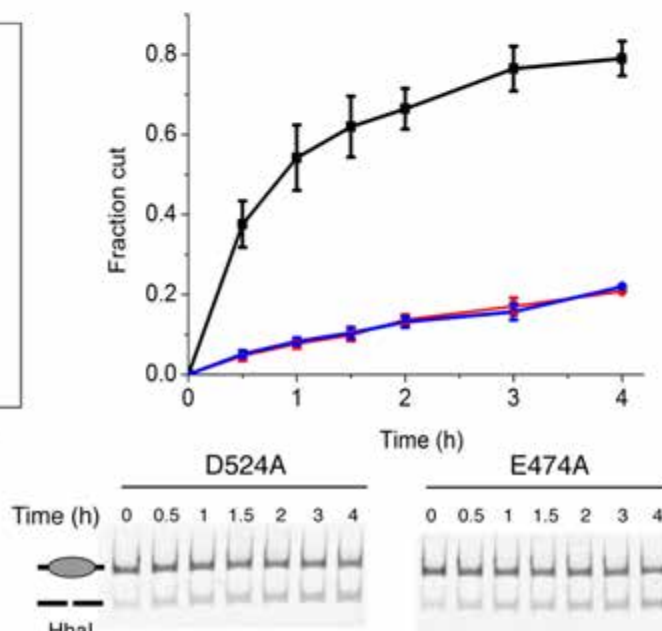
a, Representative curves of the MESG-based assays to measure the ATPase activities of MtlSWI (81–723) with intact interface (Core; black), R149A/R151A (2RA; blue), and R141A/R149A/R151A (3RA; red). The assays were performed in the absence (open circle) and presence (filled circle) of DNA. The rates of ATP hydrolysis were extracted from the slopes of the curves in the linear ranges. The activities were normalized to the ATPase activity of the Core protein in the presence of DNA. The right panel shows the quantification of the measurements in the presence (filled bars) and absence (open bars) of DNA. Error bars, s.d. ($n = 3$). **b**, Gels of the restriction-enzyme-accessibility assays of MtlSWI (81–723) with the intact interface (Core) and five L3 loop mutants. The assays were performed

with 3 nM Cy5-labelled mononucleosomes, and 5 μ M of various ISWI proteins at the indicated time points. Owing to the very low activity of the enzymes, a large excess of the proteins was used. **c**, Quantification of the remodelling activities in **b**. Core, black; R151A, green; R149A, cyan; R141A, brown; 2RA, blue; 3RA, red. Error bars, s.d. ($n = 3$). **d**, Gels of the restriction-enzyme-accessibility assays of MtlSWI (81–723) with the intact interface (Core), Δ L3 loop, and $\Delta\alpha$ 4 mutants. Quantification of the remodelling activities is showed in Fig. 2f. The assays were performed with 3 nM mononucleosomes and 0.2 μ M of various ISWI proteins. **e**, Gels of the restriction-enzyme-accessibility assays of core2 mutants of MtlSWI (81–723). The cut fractions were quantified and shown in Fig. 2i. Three independent assays were performed and one is shown.



Extended Data Figure 5 | Analyses of the interaction between the H4 tail peptide and ISWI. a, Superimposition of the final structure around the H4 peptide (yellow) with the omit difference map (grey, $F_o - F_c$, contour level $\sigma = 2$) before the H4 peptide was modelled into the structure.

b–e, ITC analyses of the interactions between the core2 domain of MtISWI and various H4 peptides. **b**, Wild-type unmodified H4 peptide; **c**, R17A mutant H4 peptide; **d**, acetylated H4K16 peptide; **e**, R19A mutant H4 peptide.

a**b****c****d****Extended Data Figure 6 | Analyses of the H4-binding surface of ISWI.**

a, GST pull-down assays. GST-H4 pulled down a significant amount of the 2D mutant MtlSWI (81–723) (lane 4). Introduction of additional D520A mutation showed a mild defect in H4-binding (lane 7). The mutations of D524A and E474A dramatically reduced the binding (lanes 10 and 13, respectively). **b**, Gels of the restriction-enzyme-accessibility assays of the H4-binding interface mutants. The assays were performed with 3 nM Cy5-labelled mononucleosomes and 0.2 μ M ISWI proteins.

The cut fractions were quantified and shown in Fig. 3e. **c**, ATPase activities of MtlSWI (81–723) with the intact interface (Core) and two mutants of the H4-binding surface (E474A and D524A) in the absence (open bars) and the presence (filled bars) of DNA. Error bars, s.d. ($n=3$). **d**, Chromatin remodelling activities of Core (black), E474A (blue) and D524A (red). The bottom panels show the representative gels of the chromatin remodelling assays. The assays were performed with 3 nM Cy5-labelled mononucleosomes and 5 μ M ISWI proteins. Error bars, s.d. ($n=3$).

Extended Data Table 1 | Data collection and refinement statistics (molecular replacement)

	ISWI ₈₁₋₇₂₃	ISWI ₄₀₆₋₇₅₄ - H4
Data collection		
Space group	P3 ₁	P1
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	127.767, 127.767, 106.666	102.238, 119.314, 132.490
α , β , γ (°)	90.000, 90.000, 120.000	89.595, 105.959, 93.078
Resolution (Å)	50-2.40 (2.49-2.40) *	50-3.00 (3.11-3.00)
<i>R</i> _{sym} or <i>R</i> _{merge}	0.066 (0.716)	0.099 (0.526)
<i>I</i> / σ <i>I</i>	27.0 (1.9)	7.2 (1.0)
Completeness (%)	99.1 (98.0)	95.3 (94.6)
Redundancy	5.2 (3.2)	1.8 (1.7)
CC1/2	0.58	0.50
Refinement		
Resolution (Å)	32.91-2.40	42.46-3.01
No. reflections	75016	112290
<i>R</i> _{work} / <i>R</i> _{free}	19.5 / 22.6	21.8 / 27.6
No. atoms		
Protein	10099	40399
Ligand/ion	1	/
Water	355	/
<i>B</i> -factors		
Protein	56	72
Ligand/ion	39	/
Water	49	/
R.m.s. deviations		
Bond lengths (Å)	0.004	0.003
Bond angles (°)	0.800	0.583

*Values in parentheses are for highest-resolution shell.

CORRIGENDUM

doi:10.1038/nature20143

Corrigendum: Synergistic, ultrafast mass storage and removal in artificial mixed conductors

Chia-Chin Chen, Lijun Fu & Joachim Maier

Nature **536**, 159–164 (2016); doi:10.1038/nature19078

In the Abstract of this Letter, “upper limit of the relaxation time” was, in the course of condensing the text, misleadingly shortened to “upper limit”. In addition, in the Supplementary Information (on page 20) “Ag⁺ transport is not possible” should have read “Ag transport is not possible”. Also in the Supplementary Information (page 26), the sentence “The chemical resistance ... are required” was fragmented. It should have read: “The chemical resistance (R^{δ}) determining the steady-state flux at the given driving force is the simpler quantity and is determined by $\frac{1}{\sigma_{\text{ion}}} + \frac{1}{\sigma_{\text{eon}}}$, expressing the fact that both non-zero ionic (σ_{ion}) and electronic (σ_{eon}) conductivities are required”. These errors have been corrected online.

ERRATUM

doi:10.1038/nature20107

Erratum: Follicular CXCR5– expressing CD8⁺ T cells curtail chronic viral infection

Ran He, Shiyue Hou, Cheng Liu, Anli Zhang, Qiang Bai, Miao Han, Yu Yang, Gang Wei, Ting Shen, Xinxin Yang, Lifan Xu, Xiangyu Chen, Yaxing Hao, Pengcheng Wang, Chuhong Zhu, Juanjuan Ou, Houjie Liang, Ting Ni, Xiaoyan Zhang, Xinyuan Zhou, Kai Deng, Yaokai Chen, Yadong Luo, Jianqing Xu, Hai Qi, Yuzhang Wu & Lilin Ye

Nature **537**, 412–428 (2016); doi:10.1038/nature19317

In this Letter, owing to errors introduced by the typesetters, ‘CXCR5⁺’ should have been ‘CXCR5[–]’ at two places in the Abstract, to read as follows: “...more potent cytotoxicity than the CXCR5[–] subset...” and “...greater therapeutic potential than the CXCR5[–] subset ...”. In addition, author Cheng Liu should have been listed as an equally contributing author, and in the Fig. 3b legend, ‘*Id*^{–/–} mice’ should have been ‘*Id2*^{–/–} mice’. In the main text, ‘me3H3k27’ should have been ‘me3H3K27’, and two occurrences of ‘CD8 T cells’ should have referred to ‘CD8⁺ T cells’. These errors have been corrected in the online versions of the paper.

CORRECTIONS & AMENDMENTS

ERRATUM

doi:10.1038/nature20163

Erratum: Kamakura replies

M. Kamakura

Nature **537**, E13 (2016); doi:10.1038/nature19350

In this Brief Communication Arising Reply, the text: “the depth of diet medium in my rearing method is 4.0–6.0 mm, or even more” should have read: “the height of larvae reared with my rearing method is 4.0–6.0 mm, or even more”. This has been corrected online.

ADDENDUM

doi:10.1038/nature20579

Addendum: REST and stress resistance in ageing and Alzheimer's disease

Lu, T. & Liviu Aron, Joseph Zullo, Ying Pan, Haeyoung Kim, Yiwen Chen, Tun-Hsiang Yang, Hyun-Min Kim, Derek Drake, X. Shirley Liu, David A. Bennett, Monica P. Colaiácovo & Bruce A. Yankner

Nature **507**, 448–454 (2014); doi:10.1038/nature13163

We have added two columns to Supplementary Table 3 of this Article to indicate the specific antibodies used for each experiment and additional information on their usage. We hope this provides a useful resource for the application of different REST antibodies. The Supplementary Information of the original Article has been updated.

CAREERS

CLIMATE CHANGE A call for a moral revolution **p.473**

BLOG Personal stories and careers counsel
<http://blogs.nature.com/naturejobs>

NATUREJOBS For the latest career listings and advice www.naturejobs.com

NIC MCPHEE



FUNDING

Word perfect

Nervous about your grant application's chance of success? Get help to make every word count.

BY AMBER DANCE

Jiri Lukas' research centre was at a crossroads four years ago. Bankrolled by the Novo Nordisk Foundation, the organization was facing a mid-term evaluation, and its funding was at risk. Lukas, executive director of the Center for Protein Research at the University of Copenhagen, wanted to apply for a grant extension, but was worried that his efforts would be wasted. It was rare at the time for foundations that award grants for biomedical research to further their support beyond one-time, limited-term funding.

A colleague told Lukas that the science in his application was strong, but that the

application itself didn't make the best case for the societal impact and unique nature of the centre. The colleague advised Lukas to consult with scientific-communication specialists at Elevate Scientific in Malmö, Sweden. "The rest was kind of a fairy tale," Lukas says. With help from Elevate, the centre won the extension.

When it comes to seeking either government or private funding, grant writers and editors are a useful resource for scientists in both academia and industry. Scientists call on them for a variety of reasons. Some simply don't have time to do it themselves. Others know that they aren't good writers, or lack a sufficient command of English. Some

are struggling to get funding. Grant writers can help with finding the right organizations to fund a project, as well as with writing the application. They can hone and focus the message, ensure consistency between sections drafted by different authors and assure adherence to strict page limits. Grant writers and editors help with everything that isn't the science, yet can still significantly affect a proposal's chance of success.

Many researchers still go it alone in preparing grant applications, but the funding landscape has changed, and scientists are now less hesitant to ask for help, says Sheila Cherry, president of Fresh Eyes Editing in Dayton, Ohio. Many funders expect applicants to ►

► seek assistance. The written guidelines from the US National Institutes of Health (NIH), for example, make that clear: “If writing is not your forte, seek help!”

There should be no shame in asking for guidance, says Anders Tunlid, a microbial ecologist at Lund University in Sweden who has reviewed grants for the European Research Council. “We need to accept that this is the way we all do it,” he says. “I don’t think that everyone has written their proposals themselves.” Colleagues may be willing to review an application’s scientific content — but they are typically too busy to spare the hours needed for fine-tuning.

“Everyone needs a little bit of help, if only to find typos,” points out David O’Keefe, senior grant writer at the Salk Institute in La Jolla, California.

The Salk offers the service for free to its researchers, but external help comes at a price: basic editing services can run from US\$500 to thousands of dollars, depending on the application. “It’s an investment, for sure,” says Stefano Goffredo, a marine ecologist at the University of Bologna in Italy. But after spending months on a proposal, he thinks it’s worth opening his wallet to get a professional polish.

Without that polish, it’s all too easy for reviewers to quickly discount an application, says Laura Hales, principal of the Isis Group, a scientific consulting and communications service in Cambridge, Massachusetts. She has served as a reviewer herself and can attest to the fact that first impressions count for everything. “You have,” she says, “one chance.”

Independent data are essentially non-existent on how professional grant-writing services affect success rates. Companies’ claims for success range from more than three times the

average rate for NIH grants to six times the average rate for the European Union’s Horizon 2020 grants. But the companies themselves concede that they can offer no guarantees. “Just because I know the formula doesn’t mean I’m going to get every one,” says Hales.

FIND YOUR MATCH

Institutions might pay for support for a junior scientist’s first few grants, says Susan Marriott, president of BioScience Writers in Houston,

“Our role is to take all the jobs that we can from the principal investigator, so that they can focus more on the research.”

Texas, but the support can be useful for mid- to later-stage-career researchers, too. Working with Elevate Scientific was a “humbling” experience, says Lukas, even as a senior scientist. The editors

identified unclear sections, improved graphics and strengthened the logic in the proposal to communicate the message more effectively.

Senior researchers in a collaboration may also use a grant editor as a project manager to ensure that all the pieces come together in a neat package by the submission deadline. It was just such a multi-investigator project that led Bruce Johnson to call in Fresh Eyes Editing. Every author tends to use their own formatting for elements such as headings and references, he notes, and editors can give the document a consistent style. “It makes it look so much more professional,” says Johnson, chief clinical research officer at the Dana Farber Cancer Institute in Boston, Massachusetts.

Editors also catch inconsistencies and redundancies in the content. For example, a large document on lung cancer does not need

to repeat in every author’s section that it’s the leading cause of cancer deaths in the United States. And one scientist might cite a statistic that 15% of people with lung cancer have a certain mutation, whereas another might write 25%. That inconsistency could cause reviewers to think that the collaborators aren’t talking to one another, Johnson says, which would not inspire a sense of confidence that the team could carry out the project together.

Grant helpers vary in the assistance they provide, and at different stages of the proposal process. Some get involved at the very start, strategizing about where to apply for funding. “It’s not only about how you write an application,” says Ram May-Ron, managing partner with the FreeMind Group in Boston. “The search starts with identifying which funding opportunity is the best one for a particular part of a research project.”

Scientists may have heard of big funding initiatives, such as Horizon 2020, but there might be other opportunities they should consider, says Eran Har-Paz, vice-president for sales at Sunrise Projects in Rosh Ha’Ayin, Israel. “We try to build a strategy, a few alternatives to submit to,” he says. “Don’t put all your eggs in one basket.”

At this level, grant helpers may reach out to programme officers, says May-Ron. For example, they might ask whether an agency has funded similar research recently, and whether they’re at all interested in doing so again. “If you go to the right place, you’re already in a better position,” he points out.

This full-scale service comes at a price, of course. Har-Paz estimates that the simplest proposal might cost a few thousand euros, with the cost escalating to €20,000 (US\$21,414) or more for elaborate

OPPORTUNITIES ABOUND

How to become a grant writer

When Laura Hales founded a biotechnology company, her first grant application was an abysmal failure. “I think I made every mistake in the book,” she recalls. But with time and resubmissions, she got the hang of it. Now she helps others to play the grantsmanship game through her communications company, the Isis Group in Cambridge, Massachusetts.

She’s not the only one; grant professionals say that business is booming. “The demand is larger than the service supplied,” says Dan Csontos, editorial director of Elevate Scientific in Malmö, Sweden. “It’s definitely a good job market if you want to get into it.”

And it’s a job market with significant perks. One advantage: “You can do it anywhere,” says David O’Keefe, a senior grant writer at the Salk Institute in La Jolla, California, who started editing while living in Indonesia.

O’Keefe also maintains a side gig of his own called pzerofive Editing.

Certain personal attributes help for wannabe grant writers, advises Eran Har-Paz, vice-president for sales at Sunrise Projects in Rosh Ha’Ayin, Israel. “You have to be a quick learner.” A good dose of self-confidence is required too, he says, as grant writers may need to exert a bit of authority to convince scientists they know the right way to pen a proposal.

Manuscript editing is a common place to start, as is working under someone else. Grant-writing courses and certificates are available, although not crucial, particularly if one has other experience.

But the main training is simply to read and write. “There are always people who would be happy to have an extra pair of eyes on an

application,” points out Cath Ennis, a project manager and grant editor in Vancouver, Canada. It is also possible to get a feel for the grants world by participating in study-review panels or working for funders.

One thing to be prepared for, advises O’Keefe, is that it gets very busy when grant deadlines roll around. “Three times a year, you’re going to have a horrible month,” he says, referring to deadlines for the US National Institutes of Health’s R01 grants, the organization’s most commonly used funding mechanism.

Nonetheless, grant writing and editing is a good option for someone who enjoys writing about science more than actually doing research, says Ennis. “It’s a great way to stay at the cutting edge of science without having to go into the lab and pipette anything.” **A.D.**

applications. That includes not only the strategizing, but also writing the majority of the application.

Some scientists already hand off much of the writing to others. Cath Ennis, a project manager and grant writer in Vancouver, Canada, might contribute an abstract, literature review, impact statement or budget, depending on the scientists' needs — but never the research plan itself. "Our role is to take all the jobs that we can from the principal investigator, so they can focus more on the research," she says.

Other grant professionals stick to editing — but that's more than just dotting i's and crossing t's. Grant editors consider content, clarity, logic and flow.

Grant professionals can be found in a variety of places: some work for a company and others as freelancers whereas some institutions have in-house specialists (see 'How to become a grant writer'). "Start talking early," advises Marriott, who is also a virologist at Baylor College of Medicine in Houston. "Even if you don't have a grant ready yet, even if you don't know what you're going to write." It's beneficial to get on an editor's calendar as early as possible, because by the time the deadline rolls around, they could have many scientists clamouring for their attention. Later on, editors may be still able to help, but in a more limited fashion, she says.

Scientists tend to look for someone with a PhD and the right technical expertise. But the match doesn't have to be exact. "I've edited grants about nuclear physics," says Ennis, whose background is in cancer biology. "I can still catch a typo when someone's put 'proton' instead of 'photon'."

Equally important, Ennis says, is to look for editors who specialize in the kind of grant one's after — say, NIH, Horizon 2020 or foundation grants. Every programme has its own requirements, and the professional should know those inside out.

With candidates in mind, the next step is to get to know them. Ask a potential editor or writer about their process, and the services they do and don't provide, advises Cherry. "It's a lot more than just, 'What's your fee and how soon can you get this done?'" she says.

Timing and costs are, nonetheless, key questions. It's best to get an estimate in advance to avoid a surprise charge later. One should also ask for a confidentiality clause in the contract.

Then, be prepared for plenty of back-and-forth. "Remember that it's a collaborative process," says Cherry. "Don't be afraid to bring up concerns and make sure you're really collaborating." ■

Amber Dance is a freelance writer in Los Angeles.

TURNING POINT

Climate guardian

Veerabhadran Ramanathan has modelled greenhouse-gas dynamics and quantified the chlorofluorocarbon (CFC) contribution to Earth's global warming. His work at the Scripps Institution of Oceanography in La Jolla, California, shows that CFC-replacing hydrofluorocarbons (HFCs) also have a potent climate-warming effect. This finding led in October to HFCs being added to the Montreal Protocol on Substances that Deplete the Ozone Layer. He has engaged for a decade with religious leaders to act on climate change.



When did you realize that science alone might not galvanize climate-change action?

Many of my colleagues and I could see that, by mid-century, we'd shoot past 2-degrees warming, yet there was no public support for the drastic actions needed to steer us away from the cliff. I was discouraged and depressed. Then I got an e-mail telling me I'd been elected to the Pontifical Academy of Sciences in Vatican City, a body of only 80 members, one-third of whom are Nobel laureates.

How did your early contact with the Vatican affect your outlook?

I initially thought the e-mail was spam. Before I got involved with the Vatican, I didn't have the foggiest notion that religion could help to combat climate change. I've since gone on record to say that global warming has to be taught in every church, synagogue, mosque and temple before we are likely to take the sort of drastic actions necessary to head it off.

Where did your involvement lead?

At a meeting hosted by the Vatican in 2011, I teamed up with Dutch Nobel laureate Paul Crutzen to focus on glaciers. That opened my eyes to the power of the Church. In the meeting's scientific report, we included a prayer to protect humanity. There was tremendous opposition, but I stood behind its inclusion. We saw the potential of mobilizing religion to help, and proposed a Vatican-hosted meeting on sustainability. This took place in 2014 under Pope Francis.

What happened after that meeting?

In a *Science* paper that followed, we pointed out that we need a moral revolution: solving climate change requires a fundamental shift in humanity's attitude towards each other and nature (P. Dasgupta and V. Ramanathan *Science* **345**, 1457–1458; 2014). Faith leaders can make such a revolution happen. After the sustainability meeting, I had two minutes to

give a summary to the Pope in the car park. I showed him that 50–60% of climate-warming pollution comes from the wealthiest people on the planet. The bottom 3 billion contribute just 5%, but will experience the worst effects of climate change. That appealed to the Pope. He asked what to do. I told him to ask people to be better stewards of the planet.

Did you get backlash for contacting religious leaders?

I was shocked — no pushback. Scientists know we need to think outside the box. It has become a moral, ethical issue.

What happened after the Pope's encyclical, or church directive, last year on the environment?

It had a huge impact on the Paris summit, in which 175 nations agreed to limit climate-change activity. A survey of people who saw the Pope during his US visit showed a statistically significant effect on how people view climate change. Pope Francis has done what he can. It's up to us to take it from here.

What does the election of Donald Trump, who won 80% of the evangelical vote, mean for climate strategy?

The US elections and the president-elect saying that the United States would withdraw from the Paris agreement hung over November's United Nations climate-change meeting. But I don't see the vote for Trump, by evangelicals or otherwise, as a vote against climate change. I think most people are protesting against economic inequality. The elections made my work with religious leaders ten times more important. We urgently need a non-political forum where we can talk about climate change. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

REFLECTIONS ON A LIFE STORY

A fresh start.

BY M. DARUSHA WEHM

The face staring back at me from the mirror is unfamiliar yet I know it's mine. It *feels* like me, even if I still fail to recognize myself sometimes. I am told this is my actual body, that this world is the real world. The life I thought I had, the world I believed to be real — those were cleverly controlled electrical impulses sent to my brain. Not real. Not me.

Yes, I asked *why*. Of course, I asked *how*.

"There are many reasons a person might be placed in simulation," my first therapist, Kris, told me. "Rehabilitation from antisocial behaviour, military training. Interrogation." I must have looked appalled. "Oh, it isn't all like that. Some people choose it themselves, like a vacation or entertainment."

"Did I?"

Kris smiled. "If it helps you to think that, perhaps that's what you should choose to believe."

"But what was it, really?"

Kris shrugged. "Why does it matter? You're here now."

It takes many sessions and another therapist before I accept that I will never know why. Two therapists later, I honestly don't care. I don't care why my body was hooked up to IVs and electrodes, why another life was created and curated, forced into my mind. I don't even care that I can never believe in reality again.

Oh, I have no doubt that my therapists are sincere, that they believe what they say, but no one will ever convince me. How could I be sure this is real when I know that the life I thought was mine was a simulation? Once you know your own eyes can deceive you, your own memory is a rewritable disk, you can never be certain. If one life can be a puppet show, any life can be. It's turtles all the way down.

"How are we feeling this week, Gil?"

I'm thankful that Samia, my latest therapist, has finally agreed to say my name with the hard 'g' — Jill or Gillian feels more wrong than the face in the mirror. The patronizing



plural, on the other hand, I can do without.

"Fine. Better."

She nods for too long, waiting for me to say more, to unburden myself. To express my feelings.

Too bad. I can wait as long as anyone.

She purses her lips, then breaks the silence. "Do you miss them?"

I don't have to ask whom she means. I had a good life in my created world. Nothing spectacular, just an ordinary existence, but it was comfortable. Fulfilling in its own unremarkable way.

"Sometimes." I sigh and decide to throw her a bone. "The other day, I saw a woman who looked like her. Not exactly the same, just something familiar about the way she carried herself, the movements of her body. It was disconcerting, unnerving." I see my reflection in the window behind Samia's desk, the short blonde hair I still can't figure out how to style, smooth scrubbed skin from my failed attempts at trying out make-up.

"How did that make you feel?"

I think about the question, like I am meant to. That's the whole point of being here, to work through how I feel about this experience. To get past it.

"I feel... blank."

Samia writes something on her tablet, nodding. She nods a lot — all the therapists

do. "It will take time before the experiences in the simulation no longer seem real, before you settle

into your true body, your true life. It's normal to feel like this world is empty for a while. It will pass."

I nod back and arrange this face into the shape of a smile. We end the session and I leave her office, feet tracing a path back to the apartment, where I will stare into the mirror.

It took months to convince me — charts and logic and explanations and unending patience. For a long time I fought them, even though I couldn't understand why they would be trying to fool me, how they could imagine I'd ever think that my life was a lie. But it was.

That was not my beautiful wife. Those were not my maddening and marvellous children. The face and body I saw in the mirror and quietly despised were never mine to hate.

The therapists think they understand what it's like to know that your entire sense of who you are isn't real. They think that it is painful, disorienting, something that needs to be overcome. They don't understand anything.

I never did see a woman who reminded me of her. It was just the kind of thing I was expected to say, the kind of thing that would make Samia make a tick on her tablet and a note on my file. *Subject is experiencing feelings of emptiness.* There was no point in explaining that this wasn't what I meant at all.

I am not the person I thought I was, and that is a joy. I am not the person anyone needs me to be, not defined by my relationship to anyone else.

I am blank. A fresh canvas, the blinking cursor.

I look at the face in the mirror knowing this may be another prison, another interrogation or training programme. It might even be real, like Samia claims. It doesn't matter. Here, I have no history, no expectations, no baggage. I am free and unencumbered.

Free to discover whom I want to become. ■

Originally from Canada, M. Darusha Wehm currently lives in Wellington, New Zealand, after spending the past several years sailing the Pacific. She's published five science-fiction novels and many short stories, and also writes mainstream fiction.

ILLUSTRATION BY JACEY